Into the Multiverse: Methods for Studying Developmental Neuroscience

Paul Alexander Bloom

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

Abstract

Into the Multiverse: Methods for Studying Developmental Neuroscience

Paul Alexander Bloom

One major challenge in developmental neuroscience research is the sheer number of choices researchers face when addressing even a single research question. Even once data collection is complete, the journey from raw data to interpretation of findings may depend on numerous decisions. To address this issue, this dissertation explores "multiverse" analysis techniques for following many analytical paths at once in the same dataset. In chapter 1, multiverses are used to examine which analyses of age-related change in amygdala-medial prefrontal cortex circuitry are robust versus sensitive to researcher decisions. Chapter 2 uses multiverse analysis to identify optimal solutions for mitigating breathing-induced artifacts in resting-state functional magnetic resonance imaging data. Chapter 3 uses a variety of model specifications to characterize simultaneous reward learning strategies in youth contingent on both visual task cues and spatial-motor information. Despite varied approaches and goals, each of the three studies highlight the benefits of conducting multiple parallel analyses for both addressing questions in developmental neuroscience and deepening understanding of the methods used to address them.

Table of Contents

List of Figures

List of Tables

Acknowledgments

To Mom, Dad, Beth, and Hannah for putting up with my grad student antics, and putting it all in perspective

To Nim, for the unwavering trust, patience, and freedom to explore

To Monica, for the statistical lifestyle

To Andrea, my lab twin

To Bridget, Michelle, Chelsea, Anna, Nicolas, Tricia, Lisa, Lior, Charlotte, Amaesha, & Syntia for your collaboration, support, and friendship in the lab

To Mariam for your guidance with the music study, and many rounds of phenomenal feedback

To Alex for the encouragement to keep running pipelines (and sometimes, permission to stop)

To Aedan, Ash, Alanna, and Yusuf for the teamwork and debugging

To Caroline & Patrick for the backwards design

To Jon, Louis, Elizabeth, Liz, Joanna, Maria, & Claudia for support in making the PhD possible

To Connor, Jess, Taylor, Bria, Alex, Justin, Henny, and Julian for the music

To Niall, Daphna, and Mike for your feedback and serving on my committee

Thank you all so very much.

Introduction

Some researchers are motivated throughout their careers by a single burning question that drives every project. My own experience as a PhD student has never been like this. Over the past several years, my advisor has described my approach to the program as "like a kid in a candy store," as I've delved deeply into projects spanning a broad range of scientific topics. While I have always viewed the breadth of my experiences and the depth to which I've been able to pursue varied research areas as a cornerstone of the PhD, I have also struggled when asked what my research focus is. At times, I have worried that committing to one "line" of research would mean letting go of other projects near and dear to me.

However, when thinking about the overarching themes linking the projects I have worked on during the PhD, a common factor has been research methodology. More specifically, my work has gravitated towards meta-scientific questions about the analysis process itself, where vast numbers of decision paths are available to researchers (Gelman & Loken, 2014). Particularly for developmental neuroscience research, where many questions focus on longitudinal trajectories or correlational study designs, there are often myriad methods for approaching ostensibly identical research questions. Often, there is no "gold standard" choice, yet such choices can drastically influence results for studies of both behavior (Orben & Przybylski, 2019) and the brain (Botvinik-Nezer et al., 2020; Bryce et al., 2021).

The work in this dissertation most generally seeks to address how researcher-driven choices during the journey from raw data to results impact developmental research studies. In particular, each of the three studies in one way or another undertake many parallel analyses of the same data in order to build meaning over a "multiverse" of possible choices (Dafflon et al.,

2020; Steegen et al., 2016) or optimize analytical decisions to meet the needs of specific research questions. The first two studies focus on such decision-making within the context of functional magnetic resonance imaging (fMRI) studies, while the third study examines behavior during a sequential reward learning task from several analytical angles.

Though the term "multiverse analysis" was originated in reference to methods tailored for inference over many possible analytical decisions (Steegen et al., 2016), such parallel analyses of the same data can serve multiple purposes. For example, researchers can construct "specification curves" displaying the estimate of interest across many possible analytic specifications, and conduct joint inference of this estimate across the entire curve (Simonsohn et al., 2020). In this sense, specification curves are akin to systematic "sensitivity checks" for a single analysis (Rohrer et al., 2017). In addition, multiverses and accompanying specification curves allow researchers to quantify how each tested decision point (i.e. 'fork' in the analytical path) impacts the estimate of interest, and which decisions are most consequential (Dragicevic et al., 2019; Liu et al., 2021; Masur, 2021).

Particularly in neuroimaging settings, multiverse analyses are not merely useful in examining variability of results across decision points, but also for systematic benchmarking of analysis strategies to optimize desired metrics (Bridgeford et al., 2020; Ciric et al., 2017a; Clayson et al., 2021; Dafflon et al., 2020; Xu et al., 2022). Particularly given that preprocessing and software choices can give rise to varying results in neuroimaging studies (Bowring et al., 2019; Li et al., 2021), multiverses are a particularly valuable tool for developing and comparing such preprocessing and analysis methods. Accordingly, recent software developments have made multiverse analyses accessible both for neuroimaging studies (Craddock et al., 2013; Esteban et al., 2019) and more generally (Liu et al., 2021; Masur, 2019).

Despite the increasing feasibility of multiverse analyses, larger multiverses are not necessarily better. Multiverses are conceptually distinct from prediction-focused techniques such as ensemble learning methods (Hastie et al., 2009; Strobl et al., 2009) or Bayesian model averaging (Raftery et al., 1997), as individual specifications typically cannot be aggregated to achieve a more precise estimate (otherwise, greater statistical precision could always be achieved by simply running more models). From a practical standpoint, larger numbers of analyses can be costly in terms of time and computing resources. Focusing on the quantity of analyses may also make careful inspection of individual analyses more difficult. In addition, recent work has highlighted that conducting multiple smaller multiverses, as opposed to one large one, may be most effective for ensuring that included specifications are reasonable (Del Giudice & Gangestad, 2021). Thus, multiverse analyses targeted towards investigating a small set of key decision points may be most valuable when there is principled reason to believe that some analysis specifications are superior to others.

In seeking to understand impacts of researcher choices in developmental neuroscience, this dissertation takes several of the multiverse approaches mentioned above. In chapter 1 (Bloom et al., 2022), specification curves serve the goal of identifying the relative robustness of age-related changes in amygdala-medial prefrontal cortex (mPFC) circuitry. In chapter 2, multiverse analyses serve the goal of optimizing respiratory correction strategies for resting-state fMRI analyses over several decision points. Chapter 3 strays furthest from typical multiverse design, but most generally uses a series of models specified to examine different aspects of youths' decision-making strategies in a two-stage reward learning paradigm. Each of these studies are summarized below.

In chapter 1, we sought to make longitudinal growth charts of task-evoked amygdala and medial prefrontal cortex (mPFC) circuitry from ages 4-22, and examine how a variety of analytical decision points impacted such growth charts. In an accelerated longitudinal cohort, we constructed Bayesian longitudinal models of age-related change in amygdala reactivity and amygdala-mPFC functional connectivity associated with fear and neutral faces. For all analyses, we conducted specification curves with varied options for inclusion criteria, nuisance regression, region-of-interest (ROI) selection, functional connectivity methodology, fMRI analysis software, and group-level model specification. Results across many such parallel analyses indicated that age-related changes in amygdala reactivity were more robust to analytical decisions than amygdala-mPFC functional connectivity, though neither of these measures were reliable. Specification curves also yielded evidence that gPPI functional connectivity measures were particularly sensitive to preprocessing decisions. Further, exploration of longitudinal model parametrization indicated that within-participant changes in amygdala reactivity with age could not be differentiated from between-participant differences. Through this work, we highlight how such specification curve analyses can help determine the robustness of developmental neuroimaging findings. We also provide an interactive data exploration website and code tutorials for conducting similar analyses.

In chapter 2, we compared many different methods for mitigating breathing-induced artifacts in functional neuroimaging studies. In particular, we addressed the problem that while informed correction of respiratory artifacts typically requires measurement of respiration itself, most neuroimaging studies do not collect such data (i.e. a respiration belt around the participant's abdomen). This is especially problematic for developmental neuroimaging studies because respiration artifacts are often trait-like and associated with age, and may bias between-participant

analyses if not mitigated (Gratton et al., 2020). Here, we built on recently developed tools to estimate participant breathing from fMRI data alone without need for collection of peripheral respiratory belt data. Then, within both an adult (Human Connectome Project) and a developmental (Nathan Kline Institute Rockland Sample) dataset, we compared many preprocessing pipelines using both the belt and estimated respiratory data for model-based correction (RETROICOR + RVT; Birn et al., 2006; Glover et al., 2000). We also compared preprocessing with no model-based correction, and with several other correction strategies that were not directly informed by the respiratory data (notch filtering the motion parameters, global signal regression, censoring, and aCompCor). Across this "multiverse" of preprocessing pipelines, we examined which strategies optimized data retention, reliability, and mitigation of residual relationships with head motion. Overall, our results highlight that model-based respiratory corrections can be done without use of peripheral respiratory belt data, yet choice of artifact mitigation strategies for resting-state fMRI necessitates making "trade-offs" based on the priority of quality control metrics.

In chapter 3, we applied several analytical strategies in parallel to develop a more thorough characterization of reward learning behaviors during middle childhood through young adulthood. Specifically, we investigated a task typically used to parse "model-free" from "model-based" reinforcement learning strategies, where learners using both strategies associate actions with rewards, but only model-based learners use cognitive maps of the task environment for prospective planning (Doll et al., 2015). We collected data among youth ages 7-14 years old (N=62) using a version of the "two-stage" task paradigm (Daw et al., 2011; Decker et al., 2016) modified to resemble a video game format. While we did not observe expected patterns of model-based or model-free decision making within this cohort, participants were nevertheless

responsive to the task structure. In addition to conducting analyses of model-based and model-free strategies at the first stage, we also explored several models of the task at the second stage. While some results were idiosyncratic to our own task, we found that participants across several cohorts (Decker et al., 2016; Nussenbaum et al., 2020; Potter et al., 2017) as well as the current paradigm displayed reward-contingent spatial-motor decision-making, even in situations where such behavior would not increase reward probability. This project highlights both how varied analytical strategies can reveal different learning strategies within the same task, as well as tendencies of children, adolescents, and young adults assign value to spatial and motor cues in environments where the outcome of their decisions are uncertain.

Overall, the three studies included in this dissertation delve "into the multiverse" of possible analytic choices within different domains. Because conducting a large number of parallel analyses is likely a step some researchers might hope to avoid within their own studies, in each of the studies included here we work to provide clear information about which choices were most impactful and which resulted in little change in the results. While these methods are often time-consuming, many of the key findings from the studies included here emerged only through taking multiple simultaneous paths from raw data to results. Thus, we demonstrate throughout that comparison of multiple analytic strategies can be broadly beneficial across a variety of research questions. To decrease barriers to entry into multiverse analyses, for each of the three studies we provide open-source code or tutorials to inform or guide readers in their own explorations.

Chapter 1: Age-related change in task-evoked amygdala—prefrontal circuitry: a multiverse approach with an accelerated longitudinal cohort aged 4-22 years

Paul Alexander Bloom, Michelle VanTieghem, Laurel Gabard-Durnam, Dylan G. Gee, Jessica Flannery, Christina Caldera, Bonnie Goff, Eva H. Telzer, Kathryn L. Humphreys, Dominic S. Fareri, Mor Shapiro, Sameah Algharazi, Niall Bolger, Mariam Aly, & Nim Tottenham

Please note, chapter published as:

Bloom, P. A., VanTieghem, M., Gabard-Durnam, L., Gee, D. G., Flannery, J., Caldera, C., Goff, B., Telzer, E. H., Humphreys, K. L., Fareri, D. S., Shapiro, M., Algharazi, S., Bolger, N., Aly, M., & Tottenham, N. (2022). Age-related change in task-evoked amygdala—prefrontal circuitry: A multiverse approach with an accelerated longitudinal cohort aged 4–22 years. *Human Brain Mapping*. https://doi.org/10.1002/hbm.25847

Abstract:

The amygdala and its connections with medial prefrontal cortex (mPFC) play central roles in the development of emotional processes. While several studies have suggested that this circuitry exhibits functional changes across the first two decades of life, findings have been mixed – perhaps resulting from differences in analytic choices across studies. Here we used multiverse analyses to examine the robustness of task-based amygdala–mPFC function findings to analytic choices within the context of an accelerated longitudinal design (4-22 years-old; N=98; 183 scans; 1-3 scans/participant). Participants, recruited from the greater Los Angeles area, completed an event-related emotional face (fear, neutral) task. Parallel analyses varying in preprocessing and modeling choices found that age-related change estimates for amygdala reactivity were more robust than task-evoked amygdala–mPFC functional connectivity to varied analytical choices. Specification curves indicated evidence for age-related decreases in amygdala reactivity to faces, though within-participant changes in amygdala reactivity could not be differentiated from between-participant differences. In contrast, amygdala—mPFC functional connectivity results varied across methods much more, and evidence for age-related change in amygdala–mPFC connectivity was not consistent. Generalized psychophysiological interaction (gPPI) measurements of connectivity were especially sensitive to whether a deconvolution step was applied. Our findings demonstrate the importance of assessing the robustness of findings to analysis choices, although the age-related changes in our current work cannot be overinterpreted given low test-retest reliability. Together, these findings highlight both the challenges in estimating developmental change in longitudinal cohorts and the value of multiverse approaches in developmental neuroimaging for assessing robustness of results.

1.1 Introduction

Rodent models and human neuroimaging have provided converging evidence for the importance of the amygdala and medial prefrontal cortex (mPFC) in the development of threat processing (Adolphs, 2008; Forbes et al., 2011), emotion regulation (Pozzi et al., 2020; Silvers et al., 2015; Sullivan & Perry, 2015), and affective learning (Moriceau & Sullivan, 2006; Pattwell et al., 2016). Characterizing growth trajectories of these regions may provide insight into neural constructions underlying emotional development (Meyer & Lee, 2019). To probe amygdala–mPFC circuitry across development, face stimuli are frequently employed because they effectively engage this circuitry while being child-appropriate (Hariri et al., 2002). Though a number of studies have examined age-related changes from childhood to young adulthood in amygdala responses and amygdala–mPFC functional connectivity (FC) associated with emotional face stimuli, findings have varied widely (likely due in part to differences in sample composition and task design; see Appendix A Table 1 for details). Several studies have found age-related change in amygdala reactivity, including decreases as a function of age in response to emotional faces (Gee et al., 2013; Guyer et al., 2008; Killgore et al., 2001; Passarotti et al., 2009; Swartz et al., 2014; Telzer et al., 2015) as well as other images (Decety et al., 2012; Silvers et al., 2017b; Vink et al., 2014), increases in amygdala reactivity with age (Joseph et al., 2015a; Todd et al., 2011), developmental sex differences (Xu et al., 2021) or peaks during adolescence (Hare et al., 2008a; Vijayakumar et al., 2019). Others have observed no age-related changes (Kujawa et al., 2016; Pfeifer et al., 2011; Pine et al., 2001; Wu et al., 2016; Yurgelun-Todd & Killgore, 2006a; Zhang et al., 2019).

With task-evoked amygdala–mPFC FC, several studies have found age-related decreases from childhood to young adulthood (Gee et al., 2013; Kujawa et al., 2016; Silvers et

al., 2017a; Wu et al., 2016), while others have found increases (Decety et al., 2012; Perlman & Pelphrey, 2011; Vink et al., 2014), developmental sex differences (Xu et al., 2021), or little age-related change (Zhang et al., 2019). While some investigations have found differing age-related change for faces displaying different emotions (Killgore & Yurgelun-Todd, 2007a; Swartz et al., 2014; Vijayakumar et al., 2019), even investigations of fearful faces specifically have varied in their developmental findings for both amygdala reactivity and amygdala—mPFC functional connectivity (Forbes et al., 2011; Gee et al., 2013; Killgore et al., 2001; Wu et al., 2016, 2016; Zhang et al., 2019).

While the small sample sizes examined in many studies on amygdala–mPFC development likely contribute to differences in findings (Marek et al., 2020), especially given low reliability of many amygdala—mPFC measures (Elliott et al., 2020; Herting et al., 2017; Sauder et al., 2013), important methodological differences also exist across studies. Differences in age range or sample demographics, stimuli, task (e.g. passive viewing vs. emotion labeling or matching (Lieberman et al., 2007), task design (blocked vs. event-related; Sergerie et al., 2008), or contrast used (faces > shapes vs. faces > baseline) may also contribute to discrepancies (see Appendix A Table 1). The brain regions under investigation also differ across studies; for example, prefrontal clusters with which amygdala connectivity was assessed. Interpreting discrepancies across studies without appreciation for these methodological differences would be inappropriate, and in fact, incorrect. Yet, such differences do not account for all discrepancies in findings across studies. Variation in processing pipelines is another source of differences across studies, as varying analytic decisions can produce qualitatively different findings, even between putatively identical analyses of the same dataset (Botvinik-Nezer et al., 2020). Choices including software package (Bowring et al.,

2019), spatial smoothing (Jo et al., 2007), treatment of head motion (Achterberg & van der Meulen, 2019), parcellation (Bryce et al., 2021), and functional connectivity approach (Di et al., 2020) can also impact results and qualitatively change findings (Cisler et al., 2014). Additionally, the majority of developmental investigations of amygdala–mPFC function have studied cross-sectional samples. Because cross-sectional studies are susceptible to cohort effects and cannot measure within-participant change, longitudinal work has been recommended for better charting of developmental trajectories (Crone & Elzinga, 2015; Madhyastha et al., 2017).

Here, we used multiverse analyses to examine the robustness of developmental changes to varied analytical decisions. We focused on task-related amygdala–mPFC functional development in an accelerated longitudinal sample ranging from ages 4-22 years. We selected a task that was designed to be appropriate for young ages to characterize developmental change in amygdala–mPFC responses to fear and neutral faces across childhood and adolescence, and we asked whether findings were robust to analytical choices. This accelerated longitudinal design is an extension of the sample reported on by Gee et al. (2013). We preregistered two hypotheses (https://osf.io/8nyj7/) predicting that both amygdala reactivity (1) and amygdala–mPFC connectivity (2) as measured with generalized psychophysiological interaction models (gPPI), would decrease as a function of age during viewing of fearful faces relative to baseline (see Table 1.1 Aims 1a & 2a).

Although we did not preregister further hypotheses, we also investigated age-related changes in within-scan differences in amygdala responses across trials and FC using beta series correlations. As previous work identified associations between amygdala–mPFC FC and separation anxiety (Carpenter et al., 2015; Gee et al., 2013), we asked whether any amygdala–

mPFC measures were associated longitudinally with separation anxiety behaviors (see Table 1.1 Aim 3). We used 'multiverse' analyses and specification curves to examine the impact of analytical decisions on results. We also investigated test-retest reliability of all brain measurements across longitudinal study visits, given the importance of such reliability for interpreting individual differences or developmental change (Herting et al., 2017). Our multiverse approach allows us to thoroughly explore the robustness of different findings to analytical choices, highlighting the importance of considering both robustness and reliability in developmental research.

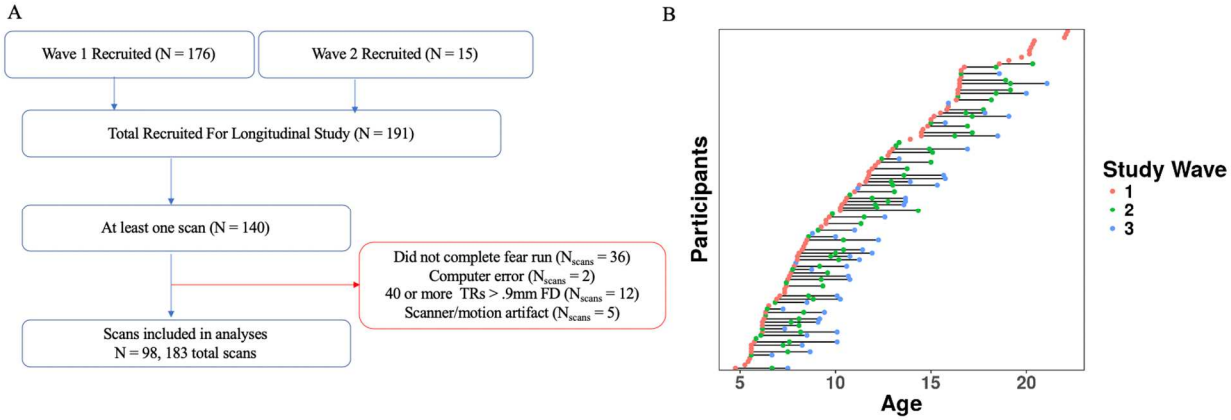


Figure 1.1: A. Schematic showing study inclusion criteria. **B.** Included scans at each study wave, with each dot representing one scan, and horizontal lines connecting participants across study waves.

Aim	Preregistered Hypothesis	Analysis Methodology	Key Findings

1a. Age-related change in amygdala reactivity to fear faces	Amygdala reactivity to fearful faces will decrease with age, such that younger children will have more positive amygdala reactivity (higher BOLD response to fear faces relative to implicit baseline) than older youth.	Multiverse amygdala ROI (anatomically-defined) analysis using multilevel linear regression at the group level. *Multiverse decision points:* Preprocessing software, GLM software, GLM nuisance regressors, amygdala ROI definition, contrast estimate type (t-stat vs. beta estimate), HRF shape, group-level model covariates, exclusion of previously analyzed scans	• Across decision points, weak but consistent negative age-related change in amygdala reactivity to fear > baseline and neutral > baseline contrasts • No consistent evidence for age-related change in fear > neutral contrast • Longitudinal models could identify consistent between-participant differences but not within-participant age-related change
1b. Age-related change in patterns of amygdala responses across task trials	None	Multiverse analysis of slopes of amygdala reactivity across trials, and amygdala reactivity in each half of trials using multilevel linear regression at the group level, single trial models *Multiverse decision points:* Global signal subtraction, amygdala ROI definition, group-level model covariates	• On average, amygdala reactivity decreased across trials (for both fear and neutral faces) • Amygdala reactivity for earlier trials was higher at younger ages • Age-related change in amygdala reactivity to fear faces in the first half of trials, but not the second half • Similar, but somewhat weaker age-related change for neutral faces
2a. Age-related change in amygdala–mPFC functional connectivity (FC) to fear faces, as measured by generalized psychophysiological interaction (gPPI)	Amygdala–mPFC FC will decrease as a function of age such that as age increases, the valence of FC will shift from positive to negative.	Multiverse gPPI analysis with anatomically defined bilateral amygdala seed and mPFC target ROIs using multilevel linear regression at the group level. *Multiverse decision points:* Deconvolution step, mPFC ROI definition, contrast estimate type (t-stat vs. beta estimate), group-level model covariates	• No consistent evidence for age-related change in gPPI for any contrast • gPPI estimates extremely sensitive to deconvolution step in creation of regressors
2b. Age-related change in amygdala–mPFC functional connectivity to fear faces, as measured by beta-series correlation (BSC)	None for BSC specifically	Multiverse BSC analysis between amygdala and mPFC using multilevel linear regression at the group level. *Multiverse decision points:* Global signal subtraction, amygdala ROI definition, mPFC ROI definition, group-level model covariates	• No consistent evidence for age-related change in BSC for any condition • Amygdala–mPFC BSC was most sensitive to selection of mPFC ROI • Global signal subtraction reduced average amygdala–mPFC BSC, but impacts on age-related changes were small • BSC estimates were not strongly associated with gPPI estimates
3. Associations of amygdala reactivity, change in amygdala reactivity across trials, or amygdala–mPFC FC with separation anxiety	None	Multiverse multilevel linear regressions with brain measures as predictors for separation anxiety behaviors, controlling for age *Multiverse decision points:* Separation anxiety measure, FC measure, mPFC ROI (FC only), amygdala ROI, contrast, deconvolution step (gPPI only)	• No evidence that amygdala reactivity, amygdala–mPFC connectivity, or change in amygdala reactivity across trials were associated with separation anxiety behaviors

Table 1.1: Summary of main aims, hypotheses, methods, and findings

1.2 Methods

Before completing analyses, we preregistered methods for the current study through the Open Science Framework at https://osf.io/8nyj7/. Only analyses for age-related changes in amygdala reactivity and amygdala–mPFC gPPI were preregistered in detail (see Table 1.1 Aims 1a & 2a), and we did not preregister multiverse analyses. Methods detailed below include both information included in the preregistration and additional information and analyses not preregistered. Analysis code & documentation can be found at https://github.com/pab2163/amygdala_mpfc_multiverse.

Participants: Participants were recruited as part of a larger study examining brain development as a function of early life caregiving experiences. The current sample (N=98; 55 female, 43 male) included typically developing children, adolescents, and young adults covering the ages 4-22 years-old (M = 11.9 years old) who enrolled to participate in a study on emotional development. All participants were reported to be physically and psychiatrically healthy (no medical or psychiatric disorders), as indicated by a telephone screening before participation, and free of MRI contraindications. All except 4 participants fell below clinical cutoffs (see Appendix A Fig. 2) on the Child Behavior Checklist (CBCL) Total Problems, Internalizing Problems, and Externalizing Problems scales (Achenbach, 1991). The larger study also included youths with a history of institutional and/or foster care outside of the United States, who are not included here. Participants from the greater Los Angeles area were recruited through flyers, state birth records, community events, online advertising, lab website and newsletters, psychologists' offices, psychology courses at a local university (participants ages 18-22 years old only), and word-of-mouth. Each participant completed up to 3 MRI scans spaced at an average interval of 18 months between visits. Parents provided written consent, children 7+ years old gave written assent, and

children under 7 years old gave verbal assent. Study protocols were approved by the local university institutional review board. These data were collected between 2009 and 2015.

An accelerated longitudinal design was used such that participants' starting ages at scan 1 comprised the entire range of sample ages (4-22 years old), and coverage was approximately balanced across the entire age range (see Fig. 1.1B). The design was structured into 3 study 'waves' corresponding with recruitment efforts and visit protocols. Participants were overenrolled at wave 1 to account for anticipated attrition (e.g., braces, relocation, etc) to achieve the desired sample size across the three waves. While most participants were recruited such that their first scan occurred at wave 1, a smaller set of participants were recruited at wave 2, such that some participants completed their first scan at wave 2 (see Fig. 1.1). For these participants, only 2 scans were planned.

Of the 191 participants participating in the longitudinal study, 140 completed at least one MRI scan. After exclusions for incomplete task runs (including falling asleep), computer errors resulting in missing stimulus timing files, high head motion, and failed visual QA (scanner/motion artifacts), a final sample of 98 participants (N = 183 total scans) was included for analysis (see Fig. 1.1). Exclusion criteria were preregistered after conducting preliminary preprocessing, but before construction of group-level models and multiverse analysis plans. This sample included 40 participants with 1 scan, 31 with 2 scans, and 27 with 3 scans (one more participant than preregistered due to an initial coding error). Wave 1 data from forty-two of these participants were reported on by Gee et al. (2013).

The median annual household income for participating families was $85,001-$100,000 (for reference, median annual household income in Los Angeles County from 2015-2019 was $68,044; US Census Bureau, 2021). Epidemiological methods were not used to recruit a sample

representative of the Los Angeles or United States populations (Heeringa et al., 2004), and Hispanic or Latinx participants were particularly underrepresented. Further sample demographics can be found in the supplementary materials (see Appendix A Tables 2-3, Appendix A Figs. 1-2).

Separation Anxiety: For each participant (except for 10 adults 18-22 years), a parent completed both the Revised Children's Anxiety and Depression Scale (RCADS-P) and the Screen for Child Anxiety Related Emotional Disorders (SCARED-P) to assess the frequency of symptoms of anxiety and low mood (Birmaher et al., 1999; Chorpita et al., 2000). Following prior work suggesting associations between task-evoked amygdala–mPFC functional connectivity and separation anxiety (Carpenter et al., 2015; Gee et al., 2013), we used the separation anxiety subscales from both the SCARED-P and RCADS-P as measures of anxiety-related behaviors in asking whether such functional connectivity may be linked to anxiety levels during childhood and adolescence. For 11 participants who had missing items on the SCARED-P, indicating parents had skipped or forgotten to answer a question, we imputed responses using 5-Nearest Neighbor imputation using only the other items included in the SCARED-P separation anxiety subscale (Beretta & Santaniello, 2016). As expected, raw separation anxiety scores on both measures decreased as a function of age, while standardized scores (which are normed based on gender and grade level) were consistent across development with few children at or near clinical threshold (see Fig. 1.6).

Emotion Discrimination Task: Participants completed either two (at wave 3) or three (at waves 1 and 2) runs of a modified 'go/no-go' task with emotional faces during fMRI scanning. Runs varied by emotional expression (fear, happy, sad), and within each run participants viewed emotional faces interspersed with neutral faces. To ensure that participants

were paying attention, they were asked to press a button whenever they saw a neutral face (no response was required for any other face expression). The order of the runs was counterbalanced across participants; the stimuli within each run were pseudorandomized (Wager & Nichols, 2003) to allow for event-related estimates of the hemodynamic response, and fixed across participants. For the present analysis, only the fear run of the task was used. The other two runs, which used happy and sad faces in place of fear, are not included in the present analysis as these conditions were not present at all waves of data collection. As 50% of trials were 'go' trials under this paradigm, we refer to the task as an emotion discrimination task, rather than a true 'go/no-go' paradigm since there was no strong prepotent motor response. Stimuli within each run were presented with a jittered ITI (3-10s, Median = 4.93s]) according to a genetic algorithm with a fixation cross on the screen (Wager & Nichols, 2003). Face images were adult White female faces from the Karolinska Directed Emotional Faces database (Calvo & Lundqvist, 2008), and the same face stimuli were used across longitudinal study visits (Vijayakumar et al., 2019). Each run (130 TRs, duration of 4:20) consisted of 48 trials (24 neutral faces, 24 fearful faces), each presented for 350ms. All fMRI analyses of this task used event-related designs.

MRI Acquisition: Participants under 18-years-old completed a mock scanning session before the MRI scan to acclimate to the scanner environment and practice lying still for data collection. Waves 1 and 2 were collected on a Siemens 3T TIM Trio MRI scanner using a standard radiofrequency head coil. A 2D spin echo image (TR, 4000 ms; TE, 40 ms; matrix size, 256 x 256; 4 mm thick; 0mm gap) was acquired in the oblique plane to guide slice configuration in both structural and functional scans. A whole-brain high-resolution T1-weighted anatomical scan (MPRAGE; 256 x 256 in-plane resolution; 256mm FOV; 192 x 1 mm sagittal slices) was acquired for each participant for registration of functional data. The task was presented through

MR-compatible goggles during scanning. T2*-weighted echoplanar images (interleaved slice acquisition) were collected at an oblique angle of ~30 degrees (130 volumes/run; TR=2000ms; TE=30 ms; flip angle=90°; matrix size=64 x 64; FOV=192 mm; 34 slices; 4 mm slice thickness; skip=0 mm). Wave 3 was collected on a Siemens 3T TIM Trio MRI scanner at a different location using identical acquisition parameters.

Behavioral Analyses: We used multilevel logistic regression models to estimate age-related changes in several task performance metrics. We fit separate models for the d' performance metric, overall accuracy (probability of a correct response on any trial), hit rate (on neutral face trials), and false alarm rate (on fear face trials) as the respective outcomes, and included nested random effects for task sessions within participants (models were not nested for d' as this analysis used only 1 metric per session rather than trial-wise outcomes, but still included random effects for participants). Additionally, to model age-related change in reaction times during correct hit trials, we fit linear, quadratic, cubic, and inverse age (1/age; Luna et al., 2004, 2021) regressions with identical random effects structures. Model equations and results for all behavioral analyses can be found in the supplement (see supplemental methods p.12-14, Appendix A Figs. 3-4).

Preregistered fMRI Pipeline: 3dskullstrip from the Analysis of Functional NeuroImages (AFNI, v20.1.16) software package (Cox, 1996) was first run on all MPRAGE scans. Next, experimenters checked the quality of the skull stripping. If there were outstanding issues with a particular scan run (areas of brain tissue cut off, or significant areas of skull left in, 30/195 scans), FSL's brain extraction tool (BET; Jenkinson et al., 2012) was used instead. We used robust brain center estimation, and modified the fractional intensity values between 0.5-0.7 to optimize quality. Slice-time correction was not used. Timeseries of the 6 motion parameters

were calculated and subsequent spatial realignment of BOLD volumes was completed using MCFLIRT in FSL (Jenkinson et al., 2002). Scans over a threshold of >40 volumes with > .9mm framewise displacement (framewise displacement calculated as the sum of absolute frame-to-frame differences between head realignment estimates; Power et al., 2012) were excluded from analysis (12 out of an initial 195, or 6.2%). After this exclusion, an average of 96.7% (range = [70.1-100%]) of stimulus-coincident volumes in each scan were below the 0.9mm framewise displacement threshold. The mean age of participants with excluded scans was 7.16 and 8/12 were male. Registration matrices were calculated for registration of functional images to high-resolution structural T1 images using FSL's FLIRT with boundary-based registration. Registration matrices for standard MNI space were also calculated using both FLIRT (linear registration) and FNIRT (nonlinear registration) with 12 DOF and a warp resolution of 10mm. Data were high-pass filtered at .01Hz and smoothed with an isotropic Gaussian kernel with FWHM of 6mm before running general linear models (GLMs), and 4d volumes were grand mean scaled such that the average intensity value was 10000.

Following preprocessing, we ran scan-level GLMs using FSL's FEAT (v6.00). We included event-related regressors for fear and neutral faces (convolved with a double-gamma HRF), their temporal derivatives (Pernet, 2014), and 24 head motion nuisance regressors (the 6 head realignment parameters, their temporal derivatives, and their squares (Power et al., 2012). Volumes with FD > .9mm were downweighed to 0 in the GLM. Pre-whitening was used to estimate and remove temporal autocorrelation (Woolrich et al., 2001). For each scan, we calculated fear > baseline, neutral > baseline, and fear > neutral contrasts. We used native-space bilateral amygdala masks generated using Freesurfer (v6.0; Fischl, 2012) by VanTieghem et al. (2021).

Multiverse Analyses and Specification Curves: In addition to the preregistered pipelines, we conducted multiverse analyses to address all aims in Table 1.1 and constructed sets of separate specification curves for each aim (see Table 1.2). In general, multiverse analyses aim to probe the consistency of results across all 'reasonable' possible combinations of analysis decisions (i.e. simultaneously taking all possible 'forking paths')(Steegen et al., 2016). Because analyzing fMRI data using all reasonable specifications was infeasible (i.e., possibilities are virtually infinite), we took the approach of 'sampling' from the many reasonable or commonly-used analysis choices for each multiverse. Despite not being completely comprehensive, this approach still allowed for thorough investigation into the robustness of results. For all multiverse analyses, we constructed specification curves by ranking models by their beta estimates (ascending) for parameters of interest for interpretation and visualization (Cosme & Lopez, 2020; Klapwijk et al., 2019; Orben & Przybylski, 2019; Simonsohn et al., 2015, 2020). Because specification choices were not preregistered, we did not conduct formal null hypothesis testing of specification curves. Instead, as continuous measures of evidence, we report the proportion of specifications resulting in an estimate of the same sign, as well as the proportion of specifications resulting in 95% posterior intervals excluding 0 in the same direction. In addition, to analyze in more detail the impact of specific choices, we submitted point estimates for parameters of interest across all specifications to multiple regression models. From these models, we examined the conditional effects of each analysis decision point on the parameter of interest (see supplemental methods p.30-31, Appendix A Figs. 11-13, 41-43, & 55-57).

Multiverse amygdala reactivity analyses: For amygdala reactivity analyses, we examined the robustness of age-related change estimates to a variety of analytical decisions. In addition to the preregistered FSL-based pipeline, we preprocessed data using C-PAC software

(v1.4.1; Craddock et al., 2013). We used C-PAC to take advantage of features supporting running multiple pipeline 'forks' in parallel (for example multiple nuisance regression forks using the same registration). No spatial smoothing was used in C-PAC pipelines (see supplemental methods p.14). Following C-PAC and FSL preprocessing, we examined the impact of different sets of commonly-used analysis methods on age-related change in amygdala reactivity. We varied analysis choices of GLM software, hemodynamic response function, nuisance regressors, first-level GLM estimates, amygdala ROI, exclusion criteria (exclude vs. include scans analyzed by Gee et al., 2013), group-level model outlier treatment, and group-level model covariates (see Table 1.2 & supplemental methods p.14-17). Multiverse analyses of amygdala reactivity included a total of 2808 model specifications (156 ways of defining participant-level amygdala reactivity x 18 group-level model specifications) for each contrast. We analyzed all specifications in parallel. In addition, we examined nonlinear age-related changes using quadratic and inverse age models (see Appendix A Figs. 14-17) and ran a smaller set of analyses (Appendix A Fig. 19) to ask whether we could differentiate within-participant change over time from between-participant differences through alternative model parametrization (see supplemental methods p.19).

For all specifications, individual-level amygdala reactivity estimates were submitted to a group-level multilevel regression model for estimation of age-related changes. All models allowed intercepts to vary by participant, and some specifications also allowed for varying slopes (see supplemental methods p.15 for model syntax). All models also included a scan-level covariate for head motion (mean framewise displacement [FD]; Power et al., 2012; Satterthwaite et al., 2012, 2013). Consistent with prior work, head motion was higher on average in younger children, and decreased with age (see Appendix A Fig. 5), though head motion was not

associated with amygdala reactivity estimates for most specifications (see Appendix A Fig. 26). Age-related change findings examined for the preregistered pipeline also remained consistent under more stringent exclusion thresholds based on mean framewise displacement (see Appendix A Fig. 27). Across preprocessing specifications, we also examined within-scan similarity of amygdala and whole-brain voxelwise reactivity patterns (see Appendix A Figs. 20-21) and between-scan correlations of average amygdala reactivity estimates (Appendix A Figs. 22-24).

Change in amygdala reactivity across trials: To probe whether amygdala reactivity exhibited within-scan change in an age-dependent manner, we modeled reactivity to each face trial using a Least Squares Separate method (LSS; Abdulrahman & Henson, 2016). After preprocessing, we used FEAT to fit 48 separate GLMs corresponding to each trial in each scan. A given trial was modeled with its own regressor and the remaining 47 trials were modeled with a single regressor. Each GLM also included 24 head motion nuisance regressors and had TRs with framewise displacement > .9mm downweighted to 0. BOLD data were high-pass filtered at .01Hz before the GLM. From each of the 48 GLMs, we extracted the mean amygdala beta estimates corresponding to a contrast for each single trial > baseline.

We constructed separate multiverse analyses using three different methods for measuring change in amygdala reactivity across trials. For method 1 (*slopes*), we measured rank-order correlations between trial number and trial-wise amygdala betas. For method 2 (*trial halves*), we split trials into the first half (trials 1-12) and second half (trials 13-24), and modeled age-related change in each half. For method 3 (*single-trial models*), we constructed larger multilevel models with individual trials as the unit of observation. We conducted several analysis specifications for each method (see Table 1.2 & supplemental methods p.21-23), and generated corresponding specification curves.

Multiverse amygdala–mPFC functional connectivity (FC) analyses: We applied multiverse analysis techniques towards examining age-related changes in amygdala–mPFC FC using gPPI and beta-series correlation (BSC) methods. Briefly, gPPI estimates functional connectivity by constructing an interaction term between the timecourse in a seed region of interest and a stimulus (task) regressor. Voxels whose activity is well fit by this interaction term (a psychological-physiological interaction, or PPI) are assumed to be "functionally coupled" with the seed region in a way that depends on the behavioral task (McLaren et al., 2012; O'Reilly et al., 2012). BSC offers a different way of estimating functioning connectivity, by constructing 'timeseries' of beta values (i.e., a beta series) in a condition of interest for two regions of interest, and calculating the product-moment correlation between those beta series.

We constructed separate specification curves for age-related change in gPPI and BSC for each contrast. Across gPPI specifications, we varied whether to use a deconvolution step in creating interaction regressors (Di & Biswal, 2017; Gitelman et al., 2003), as well as several other analysis decision points (see Table 1.2 & supplemental methods p.24-25). The deconvolution step applies to the preprocessed BOLD data from the seed timecourse: these data are first deconvolved to estimate the "underlying neural activity" that produced the BOLD signal (Gitelman et al., 2003), then these deconvolved signals are multiplied with the task regressor (e.g., for fear faces). Finally, this new interaction term is convolved with a hemodynamic response to produce the BOLD functional connectivity regressor of interest. Given recent work indicating that centering the task regressor before creation of the interaction term can mitigate spurious effects (Di et al., 2017), we also compared pipelines in which we centered the task regressor before deconvolution (pipelines including deconvolution in main analyses did not include this step; see Appendix A Fig. 44).

We preregistered constructing an mPFC ROI containing 120 voxels centered at the peak coordinates reported by Gee at al. (2013) for age-related change in fear > baseline gPPI (Talairach 2,32,8; or MNI 3,35,8). However, after preregistration we discovered that these peak coordinates were not at the center of the ROI reported by Gee at al. (2013), and were quite close to the corpus callosum. The 120-voxel ROI we created that was centered at this peak coordinate would have contained a high proportion of white matter voxels relative to cortical voxels (though this was not true for the mPFC ROI identified by Gee et al. (2013). To address this issue, we instead constructed three spherical ROIs with 5mm radii; the first centered at the above peak coordinates, the second shifted slightly anterior, and the third shifted slightly ventral relative to the second (see Fig. 1.4). Lastly, to examine amygdala functional connectivity with a more broadly-defined mPFC, we also used a 'large vmPFC' mask encompassing many of the areas within the ventromedial prefrontal cortex derived from Mackey & Petrides (Mackey & Petrides, 2014).

For BSC analyses, we used beta estimates from the LSS GLMs described above for analyses of within-scan change in amygdala reactivity. Across BSC specifications we varied analyses across several decision points (see Table 1.2 & supplemental methods p.26), including whether to include a correction for global signal (post-hoc distribution centering; Fox et al., 2009). We extracted mean beta estimates for amygdala and mPFC ROIs for each trial, then calculated product-moment correlations between the timeseries of betas across trials (neutral and fear separately) for both regions (Di et al., 2020). These correlation coefficients were transformed to z-scores, then submitted to group-level models.

Age-related changes in gPPI and BSC were estimated using multilevel regression models as described for the amygdala reactivity analyses. We focused primarily on linear age-related

change, but also examined quadratic and inverse age associations (see Appendix A Figs. 45-48 & 58-61). We separately examined group mean gPPI and BSC for each contrast (see Appendix A Figs. 38 & 52), as well as associations between mean framewise displacement and both FC measures across specifications (see Appendix A Figs. 49 & 62). Additionally, we examined mean estimates and age-related change in 'task-independent' FC as measured by beta weight of the 'physio' term from the seed amygdala timeseries within the gPPI model (representing baseline amygdala–mPFC functional connectivity controlling for task-induced variance; Appendix A Figs. 50-51).

Multiverse analyses of associations between amygdala–mPFC circuitry and separation anxiety behaviors: We used further multiverse analyses to ask whether amygdala reactivity, change in amygdala reactivity over the course of the task, or amygdala–mPFC FC were associated with separation anxiety behaviors. Separate specification curves were created for each brain measure type (amygdala reactivity, amygdala reactivity change across trials, amygdala–mPFC FC). All analyses used multilevel regression models with covariates for age, and specification curves included both RCADS-P and SCARED-P separation anxiety subscales as outcomes (see Table 1.2 & supplemental methods p.29-30). Because we did not have parent-reported RCADS-P or SCARED-P scores for 10 adult participants, these analyses had an N=173.

Reliability analyses: To better understand the proportion of variance in each measure explained by the grouping of observations within repeated measurements of the same participants over time, we computed Bayesian intraclass correlation (ICC) estimates through variance decomposition of the posterior predictive distributions of the multilevel regression models previously described. We implemented these through the 'performance' R package (Lüdecke et al., 2021; Nakagawa et al., 2017). Negative ICC estimates under this method are

possible, and indicate that the posterior predictive distribution has higher variance when not conditioning on random effects than when conditioning on them (likely indicating the posterior predictive variance is large, and random effects explain very little of this variance).

Model-fitting: All statistical models fit at the group level were run in the R (v 3.6.1) computing environment. In order to most accurately model age-related changes in each of our measures, we attempted to take into account both between-participants information and repeated measurements within participants over time. Unless otherwise indicated, models were estimated using Hamiltonian Markov chain Monte Carlo sampling as implemented in the Stan programming language through the brms package in R (Bürkner, 2019; Gelman et al., 2015). Unless otherwise indicated, all models used package default weakly-informative priors (student-t distributions with mean 0, scale parameter of 10 standardized units, and 3 degrees of freedom for all fixed effects), and were run with 4 chains of 2000 sampling iterations (1000 warmup) each (see supplemental methods p.18-19 and p.30 for syntax).

Interactive visualizations*:* Because static plots visualizing the model predictions for all models in each multiverse would require far more page space than available, we created web-based interactive visualization tools for exploring different model specifications and viewing the corresponding raw (participant-level) data and fitted model predictions using R and Shiny (Beeley, 2013). These visualizations can be found at https://pbloom.shinyapps.io/amygdala_mpfc_multiverse/

Deviations from preregistration: Although we largely completed the preregistered analyses, the current study includes many analyses beyond those proposed in the initial preregistration. Because the additional analyses (i.e., all multiverses) conducted here give us substantial analytical flexibility over that initially indicated by preregistration, we consider all

results here to be at least in part exploratory (rather than completely confirmatory), despite the preregistered hypotheses. Additionally, we note that BSC analyses, analyses of change in amygdala reactivity across trials, and analyses of associations between all brain measures and separation anxiety were exploratory, and conducted after we had seen the results of the preregistered reactivity and gPPI analyses. In addition, to avoid possible selection bias introduced by the analytical flexibility inherent in running many parallel analyses, we consider all analysis specifications simultaneously, emphasizing that without further methodological work, we consider all such choices in tandem as providing equal evidential value. While reliability analyses were not preregistered, they too provide key information for interpreting the current analyses.

Aim/Analysis	Decision Point	Choices
1a. Age-related change in amygdala reactivity to fear faces > baseline	Preprocessing Software	**FSL FEAT**, C-PAC
	GLM Software	**FSL FEAT**, AFNI 3dDeconvolve
	Hemodynamic Response Function	**Double Gamma**, Single Gamma
	Nuisance Regressors	**24 motion regressors**, 6 motion regressors, 18 motion regressors + WM + CSF
	Low-frequency artifact removal	**High-pass filter (.01Hz)**, Quadratic drift regressor
	First-level GLM Estimates	**Beta Estimates**, T-statistics
	Native vs. Standard MNI Space	**Native Space (Freesurfer)**, Harvard-Oxford Atlas in MNI
	Amygdala ROI	**Bilateral**, Left, Right, High Signal, Low Signal
	Inclusion of 45 previously analyzed scans	**Include**, Exclude
	Outlier treatment	**Exclude +/-3SD from mean**, Exclude +/-3SD from mean + robust regression
	Group-level model covariates	**Mean FD**, Mean FD + run, Mean FD + scanner, Mean FD + run + scanner
	Group-level model quadratic term	Yes, **No**
	Group-level model random slopes	**Yes**, No
1b. Age-related change in patterns of amygdala responses across task trials *FSL preproc & GLM, high-pass filter, 24 motion regressors, 2G HRF, beta estimates, included previously analyzed scans, and robust group-level regression*	Method of quantifying within-scan change	Slopes across trials, trials split into halves, single-trial models
	Global Signal Subtraction	Yes, No
	Amygdala ROI (all MNI space)	Bilateral, Left, Right
	Group-level model covariates	Mean FD**,** Mean FD + run, Mean FD + scanner, Mean FD + run + scanner
	Group-level model quadratic term	Yes, No
	Group-level model random slopes	Yes**,** No
2a Age-related change in amygdala–mPFC functional connectivity (FC) to	Deconvolution step	Yes, **No**
	mPFC ROI (all MNI space)	**3 different 5mm spheres**, large vmPFC mask

fear faces > baseline, as measured by (gPPI) *FSL preproc & GLM, high-pass filter, 24 motion regressors, 2G HRF, bilateral amygdala ROI in MNI space*	Outlier treatment	**Exclude +/-3SD from mean**, Exclude +/-3SD from mean + robust regression
	Inclusion of 45 previously analyzed scans	**Include**, Exclude
	Group-level model covariates	**Mean FD**, Mean FD + run, Mean FD + scanner, Mean FD + run + scanner
	Group-level model quadratic term	Yes, **No**
	Group-level model random slopes	**Yes,** No
2b. Age-related change in amygdala–mPFC functional connectivity to fear faces > baseline, as measured by (BSC) *FSL preproc & GLM, high-pass filter, 24 motion regressors, 2G HRF, beta estimates, robust group-level regression, included previously analyzed scans*	Amygdala ROI (all MNI space)	Bilateral, Left, Right
	mPFC ROI (all MNI space)	3 different 5mm spheres, large vmPFC mask
	Global Signal Subtraction	Yes, No
	Group-level model covariates	Mean FD, Mean FD + run, Mean FD + scanner, Mean FD + run + scanner
	Group-level model quadratic term	Yes, No
	Group-level model random slopes	Yes, No
3. Associations of amygdala reactivity, change in amygdala reactivity across trials, or amygdala–mPFC FC with separation anxiety *See supplemental methods p. 29 for details on included pipelines.*	Brain measure	Amygdala reactivity, amygdala reactivity slopes, amygdala–mPFC gPPI, amygdala–mPFC BSC
	Global Signal Subtraction (amygdala reactivity slopes & BSC only)	Yes, No
	Deconvolution step (gPPI only)	Yes, No
	mPFC ROI (all MNI space, gPPI & BSC only)	3 different 5mm spheres, large vmPFC mask
	Separation anxiety outcome variable	RCADS, SCARED raw scores, SCARED t-scores

Table 1.2: Summary of forking pipelines used in analyses for each aim[1]

[1]**Bolded** choices indicate those most closely matching preregistered pipelines.

1.3 Results

Age-related change in amygdala reactivity: We used multilevel regression models and specification curve analyses to examine age-related changes in amygdala reactivity to faces in an accelerated longitudinal sample ranging from ages 4-22 years (Fig. 1.2). Across specifications, we found relatively consistent evidence for negative age-related change in anatomically-defined (Harvard-Oxford atlas and Freesurfer-defined) amygdala reactivity to fear faces > baseline, such that the vast majority of analysis specifications (99.6%) estimated linear slopes at the group level that were negative in sign, and the majority (60.0%) of 95% posterior intervals about these slopes excluded 0 (Fig. 1.2A; interactive version at https://pbloom.shinyapps.io/amygdala_mpfc_multiverse/). Thus, over half of models indicated

that on average, increases in age were associated with decreases in amygdala reactivity to fear faces > baseline. Because the timepoint 1 data in the current study included the 42 scans used by Gee et al. (2013) to age-related changes in amygdala—mPFC circuitry for the fear > baseline contrast, results including these scans may have been more likely to find similar change (particularly for fear > baseline, see Appendix A Figs. 11-13, & 25). Estimated age-related change was on average weaker, though still largely negative (98.1% negative, 25.3% of posterior intervals excluding 0) when 42 previously analyzed scans (ages 4-17 years) were excluded to provide stricter independence from previously analyzed data (see Appendix A Fig. 11, Gee et al., 2013). Estimated average age-related change for the fear > baseline contrast was somewhat stronger when using a right amygdala ROI compared to the left amygdala, and when using t-stats extracted from scan-level GLMs rather than beta estimates for group-level models (see Appendix A Fig. 11).

Parallel multiverse analyses found similarly consistent age-related decreases in neutral faces > baseline amygdala reactivity (see Fig. 1.2C for an example pipeline & Appendix A Fig. 9 for specification curve), but no consistent evidence for age-related change for the fear > neutral contrast (see Appendix A Fig. 10). However, there was consistent evidence for higher reactivity for fear faces > neutral on average as well as each emotion compared to baseline (Appendix A Figs. 6-8), indicating that while the amygdala responses were robust and generally stronger for fear faces compared to neutral, such fear > neutral differences did not change with age. Across contrasts, varying the inclusion of block order or scanner covariates, inclusion of random intercepts, and use of robust regression models had little impact on age-related change estimates (see Fig. 1.2B Appendix A Figs. 6-8).

While group-level estimates of average age-related change were relatively consistent across specifications, the estimated age terms in these models could be influenced by both within-participant change and between-participant differences (King et al., 2018; Madhyastha et al., 2018). A smaller separate specification curve indicated that when models were parametrized to differentiate within-participant change and between-participant differences, average within-participant change was not consistent across specifications and could not be estimated with precision (Fig. 1.2D). In contrast, estimates of between-participant differences largely indicated negative age-related change in concurrence with our initial model parametrization. At the same time, within-participant versus between-participant terms were not reliably different from one another, indicating that models could not distinguish them despite higher precision for estimating between-participant differences (see Appendix A Fig. 19). We did not find consistent evidence for quadratic age-related changes in amygdala reactivity (see Appendix A Figs. 14-17). Inverse age models (i.e. amygdala reactivity modeled as a function of 1/age) indicated results similar to those of linear and quadratic models with most specifications for the fear > baseline and neutral > baseline (though less consistent) contrasts indicating age-related decreases (see Appendix A Fig. 18).

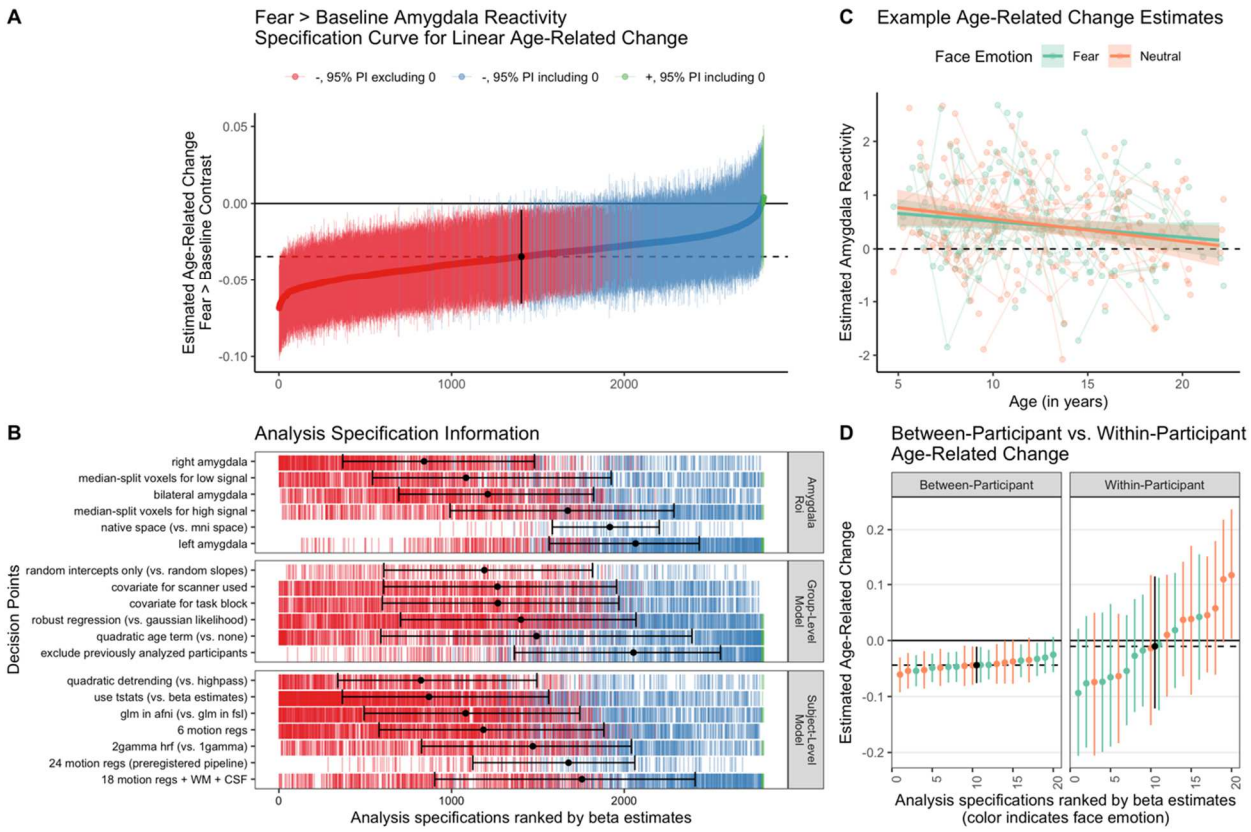


Figure 1.2: Multiverse analyses of age-related change in amygdala reactivity. **A.** Specification curve of age-related change in fear > baseline amygdala reactivity. Points represent estimated linear age-related change and lines are corresponding 95% posterior intervals (PIs). Models are ordered by age-related change estimates, with the dotted line representing the median estimate across all specifications. Color indicates sign of beta estimates and whether respective posterior intervals include 0 (red = negative excluding 0; blue = negative including 0, green = positive including 0, black = median across all specifications). **B.** Model specification information corresponding to each model in A. Variables on the y-axis represent analysis choices, corresponding color-coded marks indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis. Within each category panel (amygdala ROI, Group-Level Model, and Participant-Level Model), decision points are ordered from top to bottom by the median model rank when the corresponding choice is made (i.e. choices at the top of each panel tend to have more negative age-related change estimates). Black points with error bars represent the median and IQR ranks of specifications making the choice indicated on the corresponding line. **C.** Example participant-level data and model predictions for age-related related change in amygdala reactivity for both the fear > baseline (green) and neutral-baseline (orange) contrasts. Data are shown for a preregistered pipeline using a native space bilateral amygdala mask, 24 motion regressors, t-statistics, high-pass filtering, and participant-level GLMs in FSL. Points represent participant-level estimates, light lines connect estimates from participants with multiple study visits, and dark lines with shaded area represent model

predictions and 95% posterior intervals. **D.** Specification curves for a subset of models separately parametrizing within-participant (right) vs. between-participant (left) age-related change for both the fear > baseline (green) and neutral > baseline (orange) contrasts, as well as the median across specifications (black). See https://pbloom.shinyapps.io/amygdala_mpfc_multiverse/ for interactive visualizations.

Age-related differences in within-scan amygdala reactivity change: To ask whether age-related changes in amygdala reactivity could be due to developmental changes in patterns of amygdala reactivity across face trials (within a run), we examined whether within-scan change in amygdala reactivity varied with age (see Table 1.1 Aim 1b). Analyses included 42 specifications (3 amygdala regions of interest [ROIs] x 2 global signal correction options x 7 group-level models). Across both fear and neutral trials, linear slopes of amygdala reactivity were negative on average, indicating higher amygdala reactivity at the beginning of the run (Fig. 1.3A, Appendix A Fig. 30). Across specifications, for both fear (100% of estimates had the same sign, 95.2% of posterior intervals excluding 0 in the same direction) and neutral trials (100% of estimates in the same direction, 38.1% of posterior intervals excluding 0), there was evidence that these within-scan slopes were steeper (i.e., more negative) at younger ages, though evidence was relatively weaker for neutral trials (Fig. 1.3D-E). Specifications with a global signal subtraction step also tended to find stronger age-related change.

Similarly, when splitting trials into the first half (trials 1-12) versus second half (trials 13-24), there was consistent evidence (100% of estimates had the same sign, 69.2% with posterior interval excluding 0) for an interaction between age and trial half, such that average reactivity to fear faces > baseline in the first half of trials decreased as a function of age more so than did average reactivity during the second half of trials (see Fig. 1.3B, Appendix A Fig. 32). This interaction was in the same direction for neutral trials across most specifications (88.5% of estimates), but was typically not as strong (3.8% of posterior intervals excluding 0). Single-trial

models indicated similar age-related change in within-scan amygdala dynamics (see Fig. 1.3C, Appendix A Figs. 33-34). Mean group-level amygdala reactivity was higher for the first half of trials for fear faces > baseline across several specifications, though there were not consistent differences between trial halves for mean amygdala reactivity to neutral faces (Appendix A Fig. 31).

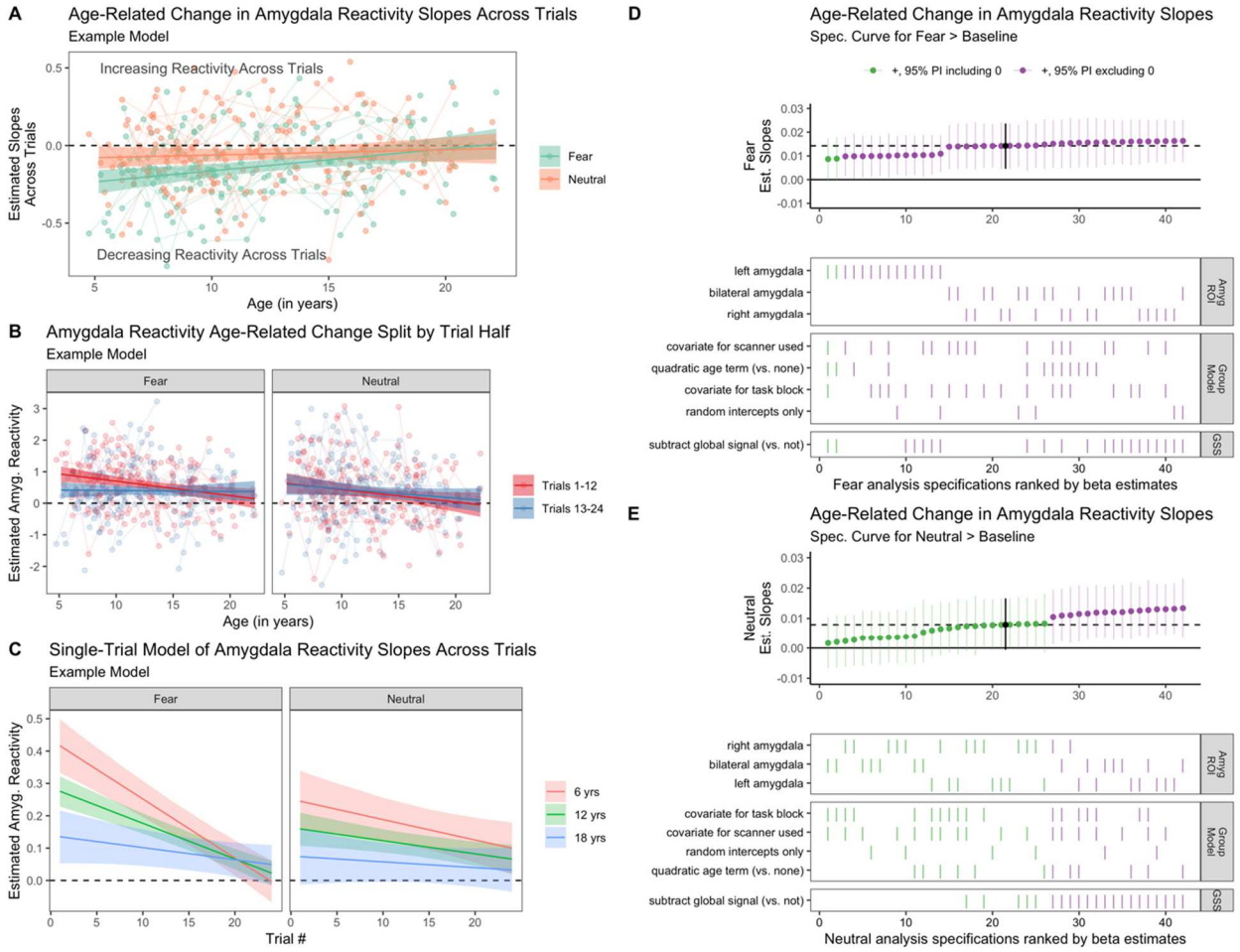


Figure 1.3: Age-related change in amygdala reactivity across trials. **A.** An example model of estimated age-related change in slopes of beta estimates across both fear (green) and neutral (orange) trials. Negative slopes represent higher amygdala activity in earlier trials relative to later trials. **B.** Example models of estimated age-related change in amygdala reactivity for the fear > baseline (left) and neutral > baseline (right) contrasts for both the first (red) and second (blue) halves of trials. In both A and B, points represent participant-level estimates, light lines connect estimates from participants with multiple study visits, and dark lines with shaded area represent

model predictions and 95% posterior intervals. **C.** Example single-trial model predictions of estimated amygdala reactivity for fear (left) and neutral (right) faces as a function of age and trial number. Age was modeled as a continuous variable, and average predictions for participants of age 6 (red), 12 (green) and 18 (blue) years are shown for visualization purposes. All estimates in A-C shown are from an example analysis pipeline using bilateral amygdala estimates and without global signal correction. **D.** Specification curve for age-related change in slopes across fear trials (i.e., many parallel analyses for the fear trials in subplot B). **E.** Specification curve for age-related change in slopes across neutral trials (i.e., neutral trials in plot B). GSS = global signal correction using post-hoc mean centering. For both D and E, color indicates sign of beta estimates and whether respective posterior intervals include 0 (green = positive including 0, purple = positive excluding 0, black = median across all specifications), and horizontal dotted lines represent median estimates across all analysis decisions. Variables on the y-axis represent analysis choices, corresponding color-coded marks indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.

Age-related change in task-evoked amygdala–mPFC functional connectivity: We used multilevel regression modelling and specification curve analyses to examine age-related change in task-evoked amygdala–mPFC functional connectivity within the accelerated longitudinal cohort (see Table 1.1 Aims 2a-b). For the fear > baseline contrast, a specification curve with 288 total specifications (4 definitions of participant-level gPPI estimates x 4 mPFC ROIs x 18 group-level models) of amygdala–mPFC gPPI did not find consistent evidence of age-related change: while 59.0% of models found point estimates in the positive direction, only 23% of posterior intervals excluded 0 (Fig. 1.4C-D, interactive version at https://pbloom.shinyapps.io/amygdala_mpfc_multiverse/). Specification curve analyses found that the sign of the estimated age-related change depended almost entirely on deconvolution, such that most specifications including a deconvolution step resulted in negative age-related change estimates never distinguishable from 0 (78.5% of point estimates negative, 0% of posterior intervals excluding 0), and most specifications not including a deconvolution step resulted in positive age-related change estimates (96.5% of point estimates positive, 47.9% of

posterior intervals excluding 0). A visualization of the effects of the deconvolution step on amygdala FC with each of four mPFC ROIs is presented in Fig. 1.4B. While mPFC ROI definition and other analysis decision points also influenced estimates of age-related change in gPPI (Fig. 1.4D), follow-up regression models indicated that the effect of including the deconvolution step was several times larger for the fear > baseline contrast (see Appendix A Figs. 41-43).

Through equivalent multiverse analyses we also found no evidence of consistent linear age-related change in amygdala–mPFC gPPI for the neutral > baseline and fear > neutral contrasts (see Appendix A Figs. 39-40), or nonlinear change for any contrast (see Appendix A Figs. 45-48). In addition, we did not see consistent evidence for group average amygdala–mPFC gPPI for any contrast, though such results often differed as a function of whether a deconvolution step was included (see Appendix A Fig. 38). Though we included gPPI analysis specifications excluding the 42 scans at timepoint 1 studied by Gee et al. (Gee et al., 2013), exclusion of these scans had little impact on age-related change results (see Fig. 4D).

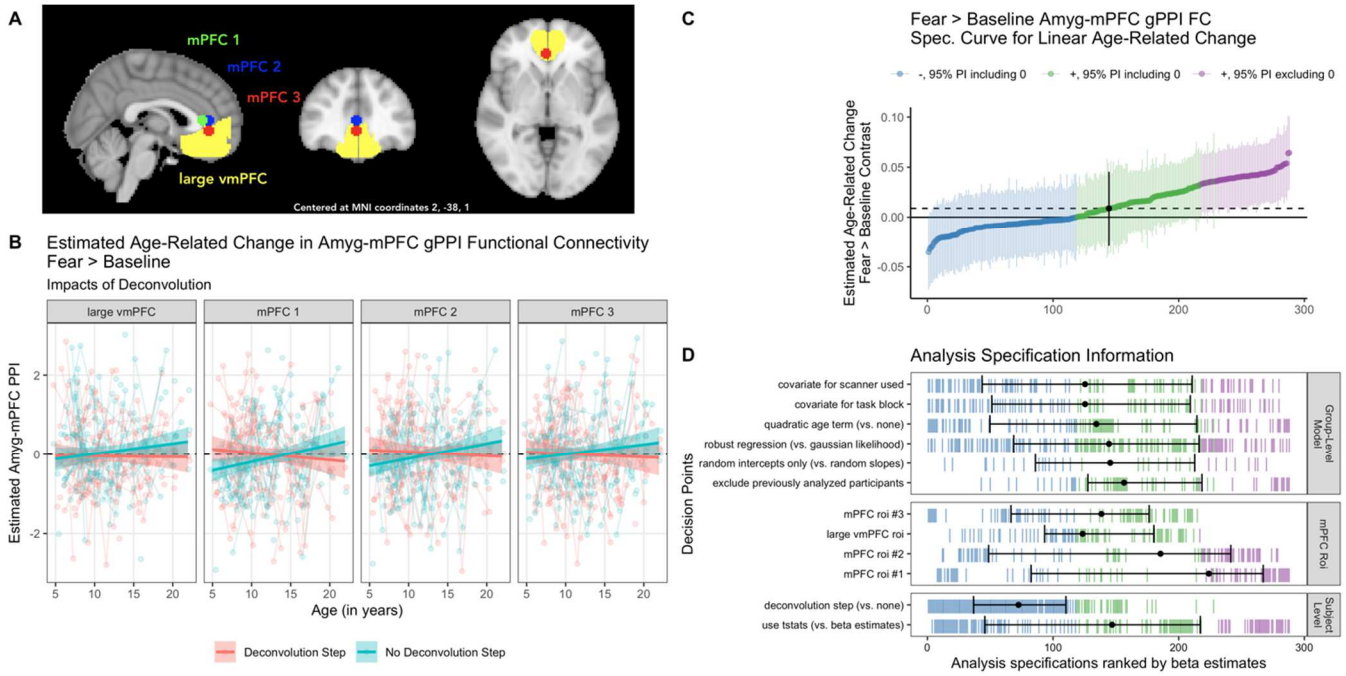


Figure 1.4: Multiverse analyses of age-related change in amygdala–mPFC connectivity using gPPI methods. **A.** MNI space mPFC ROIs used in connectivity analyses. **B.** Example participant-level data and model predictions for age-related related change in amygdala–mPFC gPPI for analysis pipelines with a deconvolution step (red), or without (blue) for each of the four regions shown in A. Although deconvolution changed the sign of age-related change estimates, the estimates are not 'statistically significant' for each pipeline alone, except for mPFC ROIs 1 & 2 without deconvolution. **C.** Specification curve of age-related change in fear > baseline amygdala–mPFC gPPI. Points represent estimated linear age-related change and lines are corresponding 95% posterior intervals. Models are ordered by age-related change estimates, and the dotted line represents the median estimate across all specifications. Color indicates sign of beta estimates and whether respective posterior intervals include 0 (blue = negative including 0, green = positive including 0, purple = positive excluding 0, black = median across all specifications). Black points with error bars represent the median and IQR ranks of specifications making the choice indicated on the corresponding line. **D.** Model specification information corresponding to each model in C. Variables on the y-axis represent analysis choices, corresponding color-coded marks indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis. Within each category (Group-Level Model, mPFC ROI, and Participant-Level Model) respectively, decision points are ordered from top to bottom by the median model rank when the corresponding choice is made (i.e., choices at the top of each panel tend to have more negative age-related change estimates). See https://pbloom.shinyapps.io/amygdala_mpfc_multiverse/ for interactive visualizations.

In addition to gPPI analyses, we used beta series correlation (BSC) analyses to examine age-related changes in task-evoked amygdala–mPFC connectivity (see Table 1.1 Aim 2b). As with gPPI, multiverse analyses of amygdala–mPFC BSC (168 total specifications; 3 amygdala ROI definitions x 4 mPFC ROI definitions x 2 global signal options x 7 group-level models) for fear trials (vs baseline) did not yield strong evidence of age-related change across pipelines (84.5% of point estimates in the same direction, 24.4% of posterior intervals excluding 0; Fig. 1.5A, interactive version at https://pbloom.shinyapps.io/amygdala_mpfc_multiverse). Unlike gPPI analyses, however, choice of mPFC ROI (as well as amygdala ROI, though this was not examined for gPPI) most impacted age-related change in BSC estimates, rather than preprocessing or modeling analytical choices (Fig. 1.5B, Appendix A Figs. 55-57). Accordingly, while global signal subtraction resulted in weaker amygdala–mPFC BSC on average (see Appendix A Fig. 52), inclusion of this step did not consistently affect age-related change estimates (Fig. 1.4C). We did not find consistent evidence for age-related change in amygdala–mPFC BSC for neutral trials (vs baseline), or for fear > neutral trials (Appendix A Figs. 53-54). We did not find consistent evidence for nonlinear age-related change for any contrast (Appendix A Figs. 58-61).

Additionally, we constructed a correlation matrix using rank-order correlations of scan-level BSC and gPPI estimates for the fear (vs baseline) condition. Across scans, there was little evidence of correspondence between BSC and gPPI metrics for amygdala–mPFC connectivity (Fig. 1.5D, Appendix A Figs. 63-66). Further, FC estimates tended to be positively correlated within a method type (BSC, gPPI) across mPFC ROIs, though less strongly for gPPI estimates with versus without a deconvolution step.

In addition to gPPI and BSC methods for functional connectivity, we also explored between-scan associations between amygdala reactivity and mPFC reactivity (Appendix A Figs. 28-29). Multilevel models indicated that amygdala reactivity for fear faces > baseline was positively associated with mPFC reactivity for fear faces > baseline for all mPFC ROIs, though we did not find consistent evidence for age-related changes in associations between amygdala and mPFC reactivity to fear faces > baseline (see Appendix A Fig. 29).

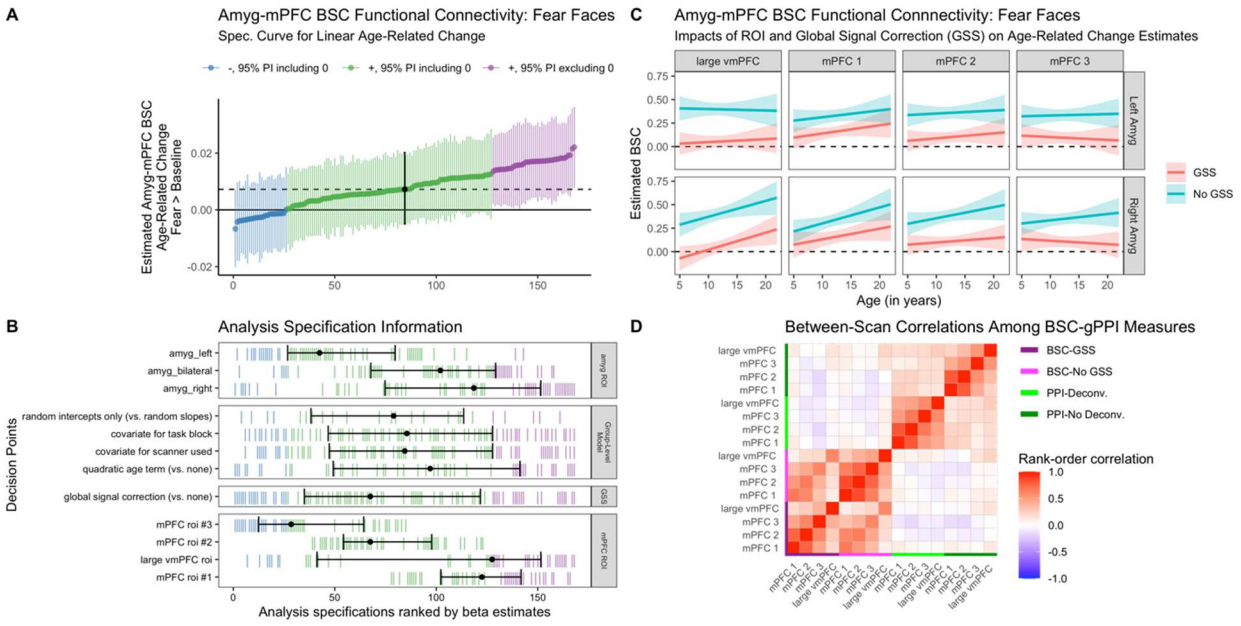


Figure 1.5: Multiverse analyses of age-related change in amygdala–mPFC connectivity using beta-series correlation (BSC) methods. **A.** Specification curve of age-related change in amygdala–mPFC BSC for fear trials. Points represent estimated linear age-related change and lines are corresponding 95% posterior intervals. Models are ordered by age-related change estimates, and the dotted line represents the median estimate across all specifications. Color indicates sign of beta estimates and whether respective posterior intervals include 0 (blue = negative including 0, green = positive including 0, purple = positive excluding 0, black = median across all specifications). **B.** Model specification information corresponding to each model in A. Variables on the y-axis represent analysis choices, corresponding color-coded marks indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis. Within each category (amygdala ROI, group-level model, global signal subtraction, and mPFC ROI) respectively, decision points are ordered from top to bottom by the median model rank when the corresponding choice is made (i.e., choices at the top of each panel tend to have more

negative age-related change estimates). Black points with error bars represent the median and IQR ranks of specifications making the choice indicated on the corresponding line. GSS = global signal correction using post-hoc mean centering. **C.** Example model predictions for age-related change in amygdala–mPFC BSC for fear trials for analysis pipelines with a global signal subtraction (GSS, post-hoc mean centering) step (red), or without (blue) for each of the four mPFC regions (see Figure 1.4A) with the left and right amygdala. Pipelines shown have random slopes, no covariates for task block or scanner, and no quadratic age term. **D.** Between-scan rank-order correlations between amygdala–mPFC connectivity measures. All gPPI measures are for the fear > baseline contrast, and BSC measures are for fear trials. See https://pbloom.shinyapps.io/amygdala_mpfc_multiverse/ for interactive visualizations.

Amygdala–mPFC Measures & Separation Anxiety: We conducted multiverse analyses of associations between several amygdala–mPFC measures (amygdala reactivity, amygdala–mPFC FC, within-scan changes in amygdala reactivity) and separation anxiety behaviors (see Table 1.1 Aim 3). Separation anxiety behaviors on average decreased with age, as indicated by the RCADS-P and SCARED-P raw scores (Fig. 1.6A-C). Neither specification curves for amygdala reactivity (18 total specifications, 56% of point estimates in the same direction as median estimate, 0% of posterior intervals excluding 0), amygdala–mPFC gPPI FC (90 total specifications, 72% of point estimates in the same direction as median estimate, 1% of posterior intervals excluding 0), amygdala–mPFC BSC FC (18 total specifications, 83% of point estimates in the same direction as median estimate, 0% of posterior intervals excluding 0), nor slope of amygdala responses across trials (12 total specifications, 75% of point estimates in the same direction as median estimate, 17% of posterior intervals excluding 0), found consistent evidence for associations between brain measures and separation anxiety. Similar specification curves found little consistent evidence for associations between brain measures and generalized anxiety, social anxiety, or total anxiety behaviors (see Appendix A Fig. 67). All specifications controlled for age (see supplemental methods p.30).

To more specifically follow up on previous work reporting associations between separation anxiety behaviors and amygdala–mPFC gPPI for fear > baseline specifically (Gee et al., 2013), we plotted model predictions for such models from the above multiverse analysis for each of the four mPFC ROIs, across all three separation anxiety outcome measures, and both with and without a deconvolution step (Fig. 1.6E). We did not find consistent evidence for associations with separation anxiety, and results showed high sensitivity to the deconvolution step, mPFC ROI, and outcome measure used.

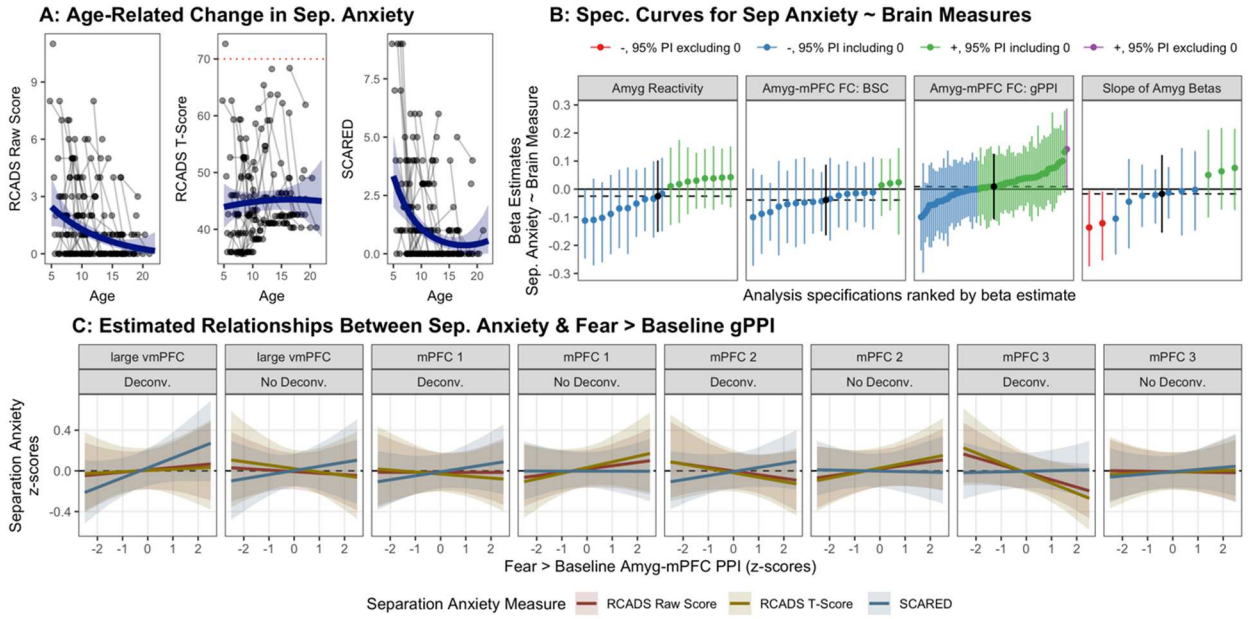


Figure 1.6: Multiverse analyses of associations between amygdala–mPFC circuitry and separation anxiety. **A.** Age-related change in SCARED and RCADS raw and t-scores for parent-reported separation anxiety subscales. The red dotted line in the middle panel represents the clinical threshold for the standardized RCADS measure (because this T-score measure is standardized based on age and gender, no age-related change is expected). **B.** Separate specification curves for associations of amygdala reactivity (left), amygdala–mPFC connectivity (both gPPI and BSC; center two panels), and amygdala reactivity slopes across trials (right) with the three separation anxiety outcomes shown in A. Points represent estimated associations between brain measures and separation anxiety (controlling for mean framewise displacement and age) and lines are corresponding 95% posterior intervals. Models are ordered by beta estimates, and the dotted line represents the median estimate across all specifications. Color

indicates sign of beta estimates and whether respective posterior intervals include 0 (red = negative excluding 0, blue = negative including 0, green = positive including 0). Scores on each separation anxiety outcome were z-scored for comparison. **C.** Example model predictions for associations between fear > baseline amygdala–mPFC gPPI and each separation anxiety measure. Predictions and 95% posterior intervals are plotted for each separation anxiety measure separately for each mPFC region, and for gPPI pipelines with and without a deconvolution step. Pipelines shown use robust regression, have random slopes, no covariates for task block or scanner, and no quadratic age term.

Reliability: To examine test-retest reliability estimates of amygdala—mPFC measures across longitudinal visits, we computed Bayesian ICC estimates using a variance decomposition method (Lüdecke et al., 2021). Because such models can accommodate missing data, all observations (98 participants, 183 total scans) were used, including participants with only 1 visit. All amygdala reactivity (Fig. 1.7A) and amygdala—mPFC functional connectivity (Fig. 1.7C) measures, as well as slopes of amygdala reactivity estimates across trials (Fig. 1.7B), demonstrated poor reliability (ICC < 0.4; Cicchetti & Sparrow, 1981; Elliott et al., 2020). Separation anxiety measures demonstrated somewhat higher, though still largely poor reliability (point estimates ~0.4, 95% CIs included values below 0.4; Fig. 1.7D). Head motion in the scanner (mean framewise displacement) showed the highest reliability (ICC = 0.52, 95% CI [0.29, 0.68]),

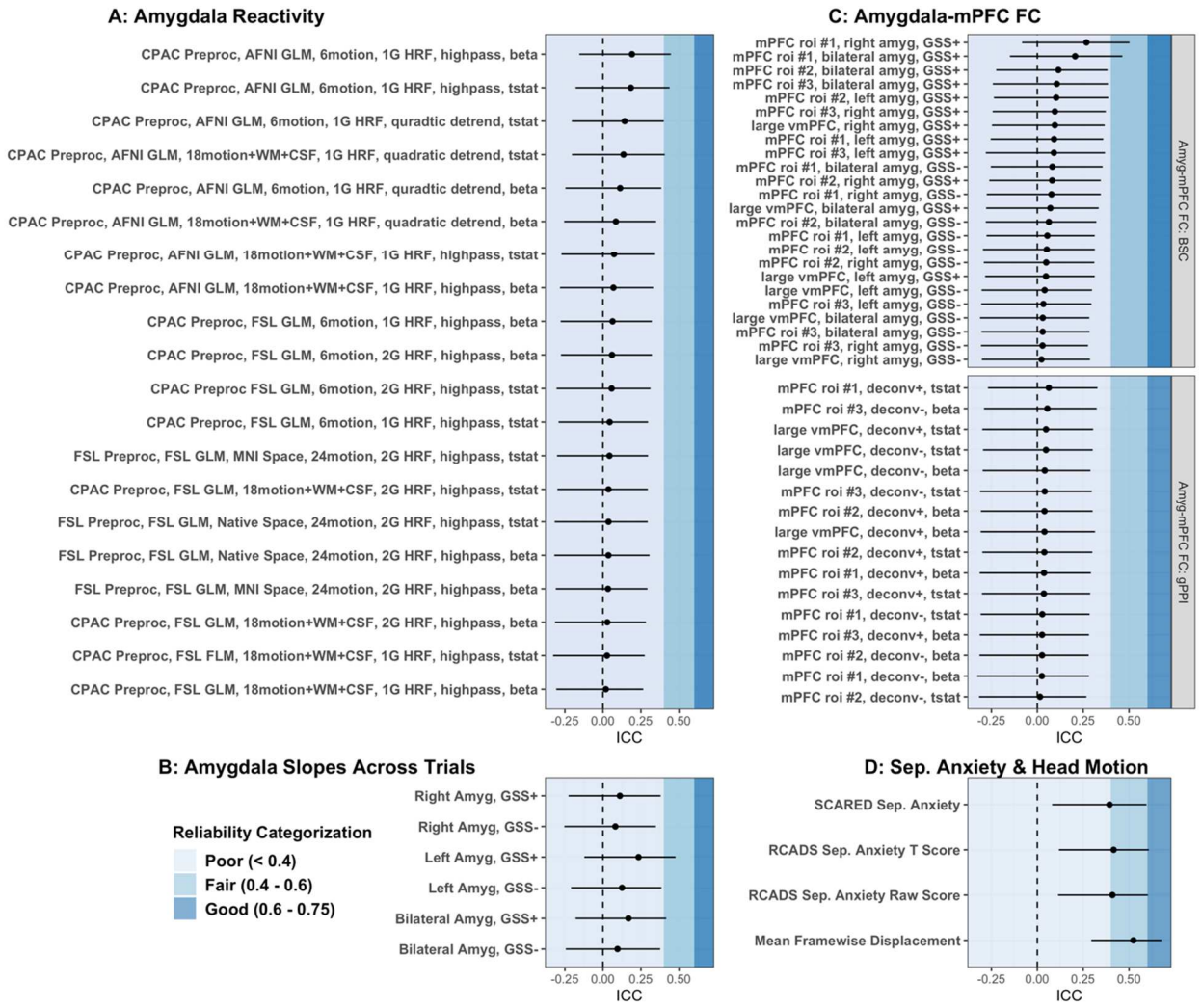


Figure 1.7: Longitudinal test-retest Bayesian ICC estimates. ICC values are shown for amygdala reactivity (A), slopes of amygdala reactivity betas across trials (B), amygdala—mPFC functional connectivity using both gPPI and BSC methods (C), and separation anxiety and in-scanner head motion measurements (D). Shaded background colors depict whether ICC estimates are categorized as poor (< 0.4), fair (0.4 - 0.6), or good (0.6 – 0.75) reliability. No ICC estimates met the threshold for excellent reliability (>0.75). Bayesian ICC estimates were calculated through a variance decomposition based on posterior predictive distributions. Negative values indicate higher posterior predictive variances not conditioned on random effect terms than conditioned on random effects terms.

1.4 Discussion

Measures that are both robust to researcher decisions and reliable across measurement instances are critical for studies of the human brain (Botvinik-Nezer et al., 2020; Bowring et al.,

2019; Elliott et al., 2020; T. Xu et al., 2022). The accelerated longitudinal design and multiverse analysis approach used in the current study allowed a rare opportunity to examine both reliability and robustness of amygdala—mPFC measures using a rapid event-related face task from early childhood through young adulthood. Overall, estimates for age-related change in amygdala reactivity were relatively robust to a variety of analytical decision points, while age-related change estimates for amygdala—mPFC connectivity were more sensitive to researcher choices. gPPI analyses were particularly sensitive to whether a deconvolution step was applied. Yet, in concurrence with previous work (Elliott et al., 2020; Haller et al., 2022; Herting et al., 2017; Infantolino et al., 2018; Kennedy et al., 2021; Nord et al., 2017; Sauder et al., 2013), amygdala—mPFC measures displayed consistently poor test-retest reliability across many analytical specifications. While low reliability estimates in the present study may be due in part to the long (~18 months) test-retest interval (Elliott et al., 2020) and potential true developmental change (Herting et al., 2017), low reliability nevertheless imposes a major caveat towards interpretation of the current developmental findings.

The present findings are valuable from a methodological standpoint in evaluating the robustness of analytical tools used. A measurement can have high test-retest reliability yet low robustness (high sensitivity) to analytical decisions, or vice versa (Li et al., 2021). Because neither robustness nor reliability guarantee the other, current findings on the impacts of analytic choices will likely be informative in guiding future studies. Thus, we discuss each of the main analyses below, with particular emphasis on how findings are impacted by analytic choices.

Amygdala reactivity: While there were differences across model specifications, the majority of pipelines supported our hypothesis that amygdala reactivity to fearful faces

decreases with age from early childhood through early adulthood (see Table 1.1 Aim 1a). Across specifications, we found relatively robust evidence for age-related decreases in amygdala reactivity to both fearful and neutral faces (Fig. 1.2A). Yet, findings also varied considerably across specifications. For example, only 60% of pipelines produced results that would be individually labeled as 'significant' (under $\alpha = .05$), indicating that multiple investigations of this dataset could likely lead to qualitatively different conclusions. While over half of analyses found evidence consistent with studies indicating greater amygdala reactivity to fear faces > baseline in younger children (Forbes et al., 2011; Gee et al., 2013; Guyer et al., 2008; Swartz et al., 2014), the other 40% of specifications would have been consistent with investigations that found little age-related change (NB: there were also differences in samples, age ranges, task parameters, and behavioral demands across these studies; Kujawa et al., 2016; Wu et al., 2016; Zhang et al., 2019). We also found that different specifications resulted in somewhat different nonlinear trajectories (see Appendix A Figs. 14-18). Not only did inverse age and quadratic age models find different trajectories (as would be expected), but quadratic trajectories themselves also displayed considerable analytic variability, with some specifications finding "convex" and others finding "concave" fits (see Appendix A Fig. 17). Although estimating nonlinear age-related change was not a primary goal of the present study, future work should use model comparisons for better differentiating nonlinear patterns (Curran et al., 2010; Luna et al., 2021).

Models also found evidence for between-participant differences, but could neither identify within-participant change (Fig. 1.2D) nor differentiate between-participant from within-participant estimates. As such, interpretation of the age-related change reported here is subject to many of the same limitations that apply to cross-sectional designs (Glenn, 2003),

where age-related changes may not necessarily indicate 'true' developmental growth. High uncertainty in estimating average within-participant change could be driven by several factors, including true heterogeneity in individual trajectories, low measurement reliability, scanner differences across longitudinal timepoints, or unmodeled variables impacting amygdala reactivity. Additionally, the within-participants terms represent a smaller age range (a maximum of 4 years for any given participant), relative to the broader age range assessed by the between-participants terms (18 years), which may have placed additional limits on identifying reliable within-participant change.

Age-related change in amygdala responses to fear faces over baseline seemed largely the result of earlier trials in the task (see Appendix A Figs. 32-34). While differences in task design and contrast across studies have been highlighted as potential sources of discrepant findings on the development of amygdala function (Killgore & Yurgelun-Todd, 2007a; Lieberman et al., 2007; Swartz et al., 2014), this result indicates that attention to trial structure and task duration may also be necessary in comparing studies. Because the paradigm used in the current study involved a task requiring participants to press for one face ('neutral') and not press for 'fear' faces, findings specific to fear faces over baseline under the current paradigm may also be driven by behavioral task demands.

Amygdala–mPFC Functional Connectivity: We did not find evidence for our second hypothesis, as neither gPPI nor BSC analyses indicated consistent evidence of age-related change in amygdala–mPFC functional connectivity (see Table 1.1 Aims 2a-2b, Fig. 1.4-5). Thus, the age-related changes in task-evoked amygdala–mPFC connectivity identified in prior work (Gee et al., 2013; Kujawa et al., 2016; Wu et al., 2016) was not identified here, consistent with (Zhang et al., 2019). Crucially, however, our specification curves did not find strong

evidence *against* such age-related change, as we did not observe precise and consistent 'null' estimates across specifications. Additionally, quadratic and inverse age models did not find consistent evidence for nonlinear age-related change (see Appendix A Figs. 45-48 & 58-61).

gPPI results were sensitive to whether a deconvolution step had been included in the preprocessing pipeline, such that we mostly found age-related decreases in amygdala–mPFC connectivity with a deconvolution step included, and age-related increases without it (although most pipelines would not have been 'statistically significant' on their own, see Fig. 1.4B). While deconvolution has been argued to be a necessary step for event-related PPI analyses (Gitelman et al., 2003), recent work has shifted guidelines on its use, and it may not be recommended for block designs (Di et al., 2020; Di & Biswal, 2017). Because the true 'neuronal' signal underlying the BOLD timeseries within a given ROI cannot be directly measured, deconvolution algorithms are difficult to validate. Further, deconvolution may cause PPI results to be driven by baseline connectivity if task regressors are not centered (Di et al., 2017), although such centering did not have a major influence on age-related change results in the present analyses (see Appendix A Fig. 44). Within the current study, small tweaks to AFNI's 3dTfitter algorithm for deconvolution resulted in vastly different regressors (see Appendix A Fig. 36), suggesting the potential for high analytic variability even between gPPI analyses ostensibly using deconvolution. While the present study does not provide evidence that can inform whether or not deconvolution is recommended, further work is needed to optimize and validate applications of gPPI methods and selection of appropriate task designs. gPPI may be better equipped for block-designs and particularly ill-posed for rapid event-related tasks due to both difficulties in resolving which times within the BOLD timeseries reflect functional connectivity evoked by rapid (350ms) events and low statistical power in

estimating such task-evoked connectivity (see Appendix A Figs. 35-37; O'Reilly et al., 2012). Concurrent with previous work, beta series correlation analyses may have higher statistical power for identifying task-related connectivity signal than gPPI within event-related designs more generally (Cisler et al., 2014).

Age-related change estimates for amygdala—mPFC BSC showed somewhat higher robustness to analytic decisions compared to gPPI. For BSC analyses, choice of mPFC ROI contributed most to variability in age-related change estimates (see Fig. 1.5B, Appendix A Figs. 55-57). While a global signal correction (post-hoc distribution centering) greatly decreased *average* amygdala—mPFC BSC connectivity (see Fig. 1.5D, Appendix A Fig. 52) for both fear and neutral faces, this analytical step did not impact age-related change estimates as heavily (Appendix A Figs. 55-57). The fact that global signal correction so dramatically decreased average estimated amygdala—mPFC BSC may indicate that, like with resting-state fMRI analyses, positive functional connectivity values are due in part to motion and physiology-related confounds (Gratton et al., 2020; Power et al., 2019). Supporting this, BSC estimates were correlated with mean framewise displacement across scans for the fear > baseline and neutral > baseline contrasts only when a global signal correction was not applied (see Appendix A Fig. 62). In addition, while test-retest reliability for all BSC measures was poor, BSC estimates from pipelines including a global signal correction step mostly demonstrated somewhat higher ICC (Fig. 1.6). While these results are consistent with prior work indicating that correcting for the global signal can mitigate artifacts (Ciric et al., 2017a; Satterthwaite et al., 2012), other work indicates that such corrections also remove meaningful biological signals (Belloy et al., 2018; Glasser et al., 2018; Yousefi et al., 2018).

Amygdala–mPFC circuitry and separation anxiety: We did not find associations between any task-related amygdala–mPFC measures (reactivity or functional connectivity) and separation anxiety behaviors (see Table 1.1 Aim 3; Fig. 1.6). This finding stands in contrast to associations between amygdala–mPFC connectivity and anxiety identified in previous developmental work (Gee et al., 2013; Jalbrzikowski et al., 2017; Kujawa et al., 2016; Qin et al., 2014). However, given that analyses of brain-behavior associations may require imaging cohorts much larger than the current sample (especially considering the low reliability of the measures used; Grady et al., 2020; Marek et al., 2020), the absence of relationships here may not be strong evidence against the existence of potential associations between amygdala–mPFC circuitry and developing anxiety-related behaviors.

Advantages and pitfalls of the multiverse approach: Our findings contribute to a body of work demonstrating that preprocessing and modeling choices can meaningfully influence results (Botvinik-Nezer et al., 2020). Indeed, most studies involving many analytical decision points could benefit from multiverse analyses. Such specification curves can help to examine the stability of findings in both exploratory and confirmatory research (Flournoy et al., 2020). Particularly when methodological 'gold standards' have not been determined, specification curves may be informative for examining the impacts of potential analysis decisions (Bridgeford et al., 2020; Dafflon et al., 2020). Further, wider use of specification curves might help to resolve discrepancies between study findings stemming from different analysis pipelines.

While specification curve analyses may benefit much future research, we also note that multiverses are only as comprehensive as the included specifications (Steegen et al., 2016), and such analyses alone do not solve problems related to unmodeled confounds, design flaws,

inadequate statistical power, circular analyses, or non-representative sampling. Further, unless all specifications are decided *a priori*, analyses are vulnerable to problems of analytic flexibility (Gelman & Loken, 2014), and inclusion of less justified specifications can bias results (Del Giudice & Gangestad, 2021). Because specification curves can include hundreds or thousands of individual analyses, rigorous evaluation of individual models can be difficult. To this end, we created interactive visualizations for visual exploration of individual analysis specifications.

Computational resources are a relevant concern when conducting multiverse analyses as well. In the current study, preprocessing (registration in particular) was the most computationally intensive step, taking an estimated 4 hours of compute time per scan per pipeline using 4 cores on a Linux-based institutional research computing cluster. However, specification curve analyses themselves were relatively less intensive, with all group-level models of amygdala reactivity completing in a total of 48 hours using 4 cores on a laboratory Linux-based server. Specification curves using maximum likelihood models (lme4 in R; Bates & Bolker, 2011) were even more efficient, with thousands of models running within minutes using a 2019 MacBook Pro (2.8 GHz Intel Core i7).

Limitations: The present study is subject to several limitations that may be addressed in future investigations. Perhaps most crucially, our conclusions (along with those of many developmental fMRI studies) are limited by the poor test-retest reliability of the fMRI data. Because amygdala—mPFC measures showed low reliability across study visits, the statistical power of our analyses of age-related changes is likely low (Elliott et al., 2020; Zuo et al., 2019). Low-powered studies can yield increased rates of both false positive and false negative results (as well as errors of the sign and magnitude of estimates; Button et al., 2013; Gelman & Carlin,

2014); therefore we caution against interpretation of our developmental findings (and brain-behavior associations) beyond the cohort studied in the present investigation. In particular, the low statistical power of our rapid event-related task design may be a major contributor to the low test-retest reliability and variance in outcomes across analysis specifications. That being said, achieving high-powered studies presents a challenge for studying populations that cannot tolerate lengthy fMRI sessions. Both findings that were more robust to analytical decisions (amygdala reactivity) and findings that were less so (amygdala—mPFC connectivity, associations with separation anxiety) may be most valuable in meta-analytic contexts where greater aggregate statistical power can be achieved. In particular, future work on amygdala—mPFC development will benefit from optimization of measures both on robustness to analytic variability (Li et al., 2021) and reliability (Kragel et al., 2021).

Present findings are also limited by the number of participants studied (Bossier et al., 2020; Marek et al., 2020), the number of longitudinal study sessions per participant (King et al., 2018), and the duration of the task (Nee, 2019). Work with larger sample sizes, more study sessions per participant, and more task data collected per session will be necessary for charting functional amygdala–mPFC development and examining heterogeneity across individuals (although collecting task-based fMRI will continue to be challenging for studies including younger children). The generalizability of the current findings may also be limited by the fact that this cohort was skewed towards high incomes and not racially or ethnically representative of the Los Angeles or United States population.

Findings are also somewhat limited by the fact that the present study is not wholly confirmatory, despite preregistration. Because our multiverse analysis approaches expanded significantly beyond the methods we preregistered, most of the present analyses, while

hypothesis-driven, must be considered exploratory (Flournoy et al., 2020). The fact that some specifications used data included in previous similar analyses of the same cohort(Gee et al., 2013) also limits the confirmatory power of the present study (Kriegeskorte et al., 2009). This may be especially true because longitudinal models could not identify within-person change as distinct from between-participant differences (see Fig. 1.2D), indicating that our age-related change estimates may be influenced by cross-sectional information similar to that investigated by Gee et al. (Gee et al., 2013).

Though the current study aimed to estimate longitudinal age-related changes in amygdala–mPFC functional circuitry evoked by fear and neutral faces, the current findings may not be specific to these stimuli (Hariri et al., 2002). Because our task did not include non-face foils or probe specific emotion-related processes, results may be driven by attention, learning, or visual processing, rather than affective or face processing. In particular, because participants were instructed to press a button for neutral faces and withhold a button press for fear faces, observed amygdala—mPFC responses may in part reflect response inhibition (for fear faces; Menon et al., 2001) and target detection processes (for neutral faces; Jonkman et al., 2003). Findings for the fear faces > baseline and neutral > baseline contrasts also may not be valence-specific in the absence of a different emotional face as part of the contrast. Further, because all faces were adult White women, the current results may not generalize to faces more broadly (Richeson et al., 2008; Telzer et al., 2012). Additionally, because face stimuli were the same across study visits, exposure effects across sessions may confound longitudinal findings (although exposure effects may be possible any time a task is repeated, even if stimuli are unique), particularly age-related decreases in amygdala responses (Telzer et al., 2018). While within-session amygdala habituation effects have been shown across several paradigms

(Geissberger et al., 2020; Hare et al., 2008a; Hein et al., 2018), between-session habituation effects are unlikely beyond 2-3 weeks (Geissberger et al., 2020; Johnstone et al., 2005; Plichta et al., 2014; Spohrs et al., 2018).

Finally, our findings on age-related change in amygdala and mPFC function may be biased or confounded by age-related differences in head motion (Ciric et al., 2017a), anatomical image quality and alignment (Gilmore et al., 2020; Rorden et al., 2012), signal dropout, and physiological artifacts (Boubela et al., 2015; Fair et al., 2020; Gratton et al., 2020). While our multiverse analyses included preprocessing and group-level modeling specifications designed to minimize some of such potential issues, future work is still needed to optimize discrimination of developmental changes of interest from such potential confounds.

Despite these limitations, the present study concords with prior investigations in demonstrating the value of multiverse approaches to quantify sensitivity to researcher decisions. The results highlight key analytic considerations for future studies of age-related changes in amygdala—mPFC function, as well as for studies of human brain development more broadly.

CRediT Author Statement:

Paul Alexander Bloom: Conceptualization, Methodology, Formal Analysis, Writing - Original Draft, Visualization **Michelle VanTieghem:** Methodology, Writing – Review & Editing **Laurel Gabard-Durnam:** Investigation, Methodology, Writing – Review & Editing, **Dylan Gee:** Investigation, Methodology, Writing – Review & Editing **Jessica Flannery:** Investigation**,** Writing – Review & Editing **Christina Caldera:** Investigation**,** Writing – Review & Editing **Bonnie Goff:** Investigation**,** Writing – Review & Editing **Eva Telzer:** Investigation**,** Writing – Review & Editing **Kathryn L. Humphreys:** Investigation**,** Writing – Review & Editing **Dominic Fareri:** Investigation**,** Writing – Review & Editing **Mor Shapiro:** Investigation**,** Writing – Review & Editing **Sameah Algharazi:** Validation**,** Writing – Review & Editing **Niall Bolger:** Methodology, Formal Analysis, Writing – Review & Editing **Mariam Aly:** Methodology, Formal Analysis, Supervision, Writing – Review & Editing **Nim Tottenham:** Conceptualization, Methodology, Investigation, Resources, Data Curation, Writing - Original Draft, Supervision, Funding Acquisition

Code Availability

Code for preprocessing, analysis, and data visualizations for this manuscript is available at https://github.com/pab2163/amygdala_mpfc_multiverse. While unfortunately this code cannot be run as written without data, we have attempted to document analysis steps clearly. In addition, we provide publicly available simulated data structured similarly to the study data on amygdala reactivity, such that interested readers can view multiverse analysis walkthroughs (https://pab2163.github.io/amygdala_mpfc_multiverse) and experiment with analysis code.

Additional materials, including MNI space masks and preregistration documentation, are available at https://osf.io/hvdmx/

Chapter 2: Addressing breathing-induced head motion in functional neuroimaging data

Paul Alexander Bloom, Yiwen Tian, Ryan Lim, Anna Vannucci, Jon Clucas, Lei Ai, Nim Tottenham, Lia Hocke, Blaise Frederick, Michael P. Milham, & Alexandre R. Franco

Abstract

Breathing-induced head 'pseudomotions' in functional MRI experiments have recently become a cause of greater concern. In addition to introducing systemic BOLD artifacts, respiration presents a problem for data quality assurance through inflation of head motion estimates. While retrospective image-based (RETROICOR) or respiration volume per time (RVT) correction can reduce artifacts, both techniques depend on respiration belt data not collected in many studies. Recently, temporal filtering of head realignment parameters with a notch filter has been proposed, though this technique also depends on knowledge of participants' breathing. Building on recent work estimating respiratory traces at high temporal resolution from BOLD data alone, we asked whether such predicted respiratory traces can correct respiratory artifacts without respiration belt data. Specifically, within the Human Connectome Project test-retest (N = 36, 24F/12M, ages 22-35) and NKI Rockland Sample (N = 97, 58F/39M, ages 6-20 years) cohorts, we compared preprocessing strategies for correcting such artifacts on data retention, reliability of functional connectomes, and residual head motion artifacts. Both notch filtering and model-based correction with predicted respiration mitigated psueodmotion. Correction using predicted respiratory traces yielded similar functional connectivity estimates to when belt traces were used, indicating viability in datasets without belt measurements. However, impacts of model-based and notch filtering correction strategies on functional connectivity estimates were minimal compared to those of GSR, aCompCor, and censoring. With the exception of aCompCor (which either showed benefits or no effects on QC metrics), most preprocessing steps involved tradeoffs between data various QC metrics. Thus, instead of a "one size fits all" approach, future studies may benefit from tailoring preprocessing strategies based on the relative priority of distinct quality assurance metrics or multiversing consequential pipeline decisions.

2.1 Introduction

Breathing-related signals are present in functional magnetic resonance imaging (fMRI) data (Chang & Glover, 2009; Prokopiou et al., 2019), particularly in studies of functional connectivity in which the measure of interest is covariance among signals (for example, "resting–state" analyses). Such respiratory artifacts pose two key issues for researchers. First, breathing can cause systematic artifacts within BOLD data that may bias estimates of functional connectivity (Birn et al., 2006; Power et al., 2020). In particular, because breathing patterns contain state and trait-like components and can vary with age (Tobe et al., 2021), sex (Lynch et al., 2020), BMI (Gratton et al., 2020), stress (Suess et al., 1980), and psychopathology (Giardino et al., 2007), such artifacts (akin to head motion artifacts) may confound relationships of interest in many fMRI studies. Second, breathing causes "pseudomotion" artifacts in BOLD data that inflate estimates of head motion (Brosch et al., 2002; Durand et al., 2001; Power et al., 2019). Such breathing-induced pseudomotion complicates detection of motion-contaminated volumes, making it more difficult to decide which data should be excluded from analysis (Fair et al., 2020). Mitigation of breathing-induced artifacts is therefore a crucial step in preparing fMRI data for analyses of functional connectivity.

While model-based methods exist for correction of respiratory artifacts in fMRI data, their applicability is limited by their reliance on peripheral physiological measurements that are often not collected during neuroimaging studies. In particular, an image□based method for retrospective correction of physiological signal in fMRI (RETROICOR; Glover et al., 2000; Tijssen et al., 2014) or Respiration Volume (RV) and Respiration Volume per Time (RVT; Birn et al., 2006, 2008, 2014a; Chang & Glover, 2009) correction can mitigate breathing-induced artifacts, yet rely on data from a pneumatic respiration belt placed around a participant's

abdomen measuring chest volume during scanning. In particular, RVT may capture signals associated with fluctuations in arterial CO2 (Chang & Glover, 2009; Golestani et al., 2015). Such tools can be applied after data collection and have shown efficacy in mitigating both global and spatially heterogeneous physiological (cardiac and respiratory) artifacts. Unfortunately, the required respiratory belt data can be difficult to set up properly, and often slip off or yield unreliable breathing data, especially among children or high-motion participants. As many fMRI studies do not acquire belt data, respiratory corrections that do not rely on such measurements are crucial for increasing the feasibility of effective artifact mitigation.

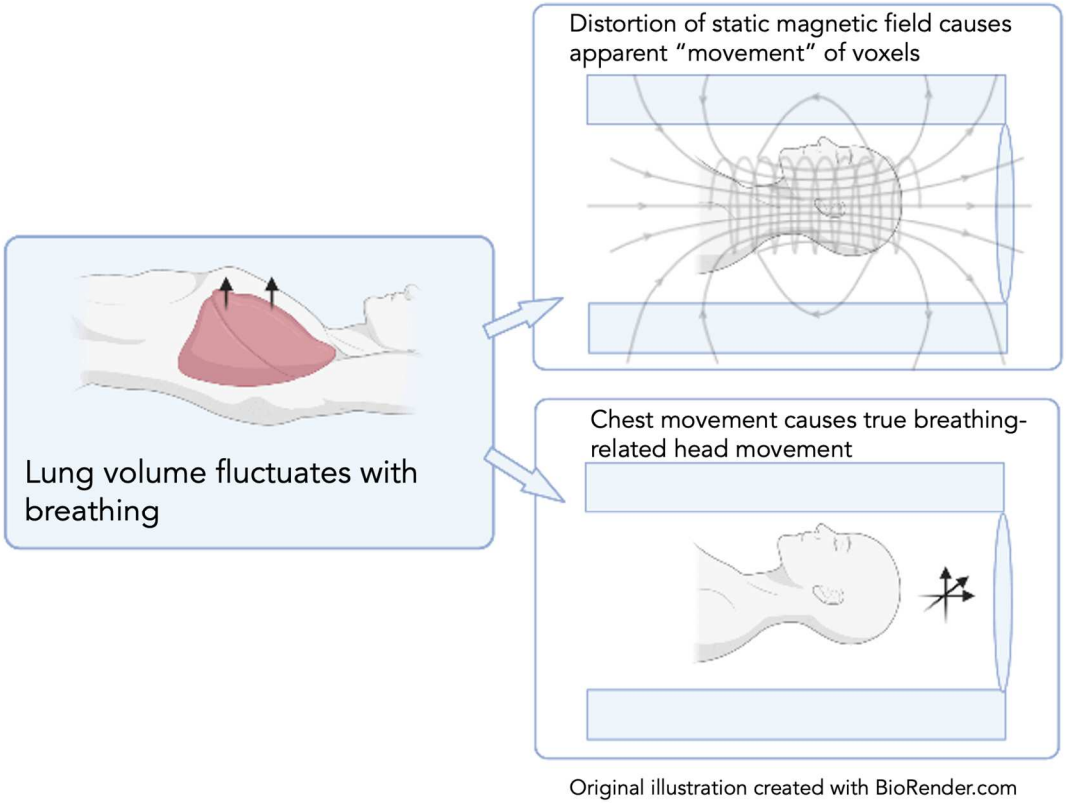


Figure 2.1. Schematic representation of simultaneous true head motion and peudomotion caused by respiration during fMRI. Specifically, BOLD images "shift" periodically in the phase-encoding direction as inhalation and exhalation occur (Raj et al., 2000, 2001). MR physics mechanisms behind such respiratory pseudomotion are beyond the scope of the current investigation (see Brosch et al., 2002).

Recently, filtering high-frequency fluctuations from the head realignment parameters has been proposed as a potential workaround for removing respiratory-induced signal from head motion calculations without requiring respiratory belt data (Fair et al., 2020; Kaplan et al., 2022). Among adults, quasi-periodic respiratory signals driven by regular breathing typically present within BOLD data at around 0.2-0.3Hz (Charlton et al., 2018), while respiration frequencies are higher on average in children (Caballero-Gaudes & Reynolds, 2017; Tobe et al., 2021). Specifically within multiband data, a notch filter (centered at the median frequency of quasi-periodic respiration across a dataset) applied to the head realignment parameters as shown success in distinguishing true from pseudo motion and saving data from being censored (Fair et al., 2020). Such filtering also improves assessment of data quality based on adjusted head motion estimates (FD) post-filtering. While single-band data sets, or those with lower temporal resolution, may not support notch filtering if the filter response range falls near or above the Nyquist folding frequency (1/2 the frequency of BOLD acquisition), recent work has demonstrated that a similar lowpass filter can improve head motion estimation and preserve single-band data from being censored (Gratton et al., 2020).

Yet, informed calibration of filters requires knowledge of respiratory frequencies within a dataset (i.e. for defining the response of a notch filter). Further, application of the same filter to the head realignment parameters of all participants in a dataset may fail to capture prominent frequencies of pseudomotion in some participants if breathing rates are heterogeneous. As breathing rates tend to decrease between childhood and young adulthood (Wallis et al., 2005), defining a common filter for datasets with wide age ranges may not be an effective approach at mitigating such artifacts. On the other hand, individual-specific filters cannot be derived without

individual-level respiratory frequency information, and low-quality respiratory belt data may preclude accurate estimation of the individual-specific filters even when available (Fair et al., 2020). In addition, even individual-specific filters may fail to capture breathing-induced psuedomotion if the frequency of individuals' respiration fluctuates within scan runs. Further, because notch filtering is applied only to head realignment estimates (for use in nuisance regression or motion-based censoring), this strategy may not on its own mitigate effects of breathing on the fluctuations in the BOLD signal itself.

Two methods that do not require peripheral respiratory belt data; regression of the global signal ("GSR") averaged across the gray matter voxels (Falahpour et al., 2013; Fox et al., 2009; Power et al., 2017), and the top principal components across atomically-defined white matter and cerebrospinal fluid voxels (anatomical component correction is often referred to as "aCompCor"; Behzadi et al., 2007; Muschelli et al., 2014); have been proposed for physiological noise correction. Indeed, global signal is thought to reflect, in part, fluctuations in arterial CO2 (Chang & Glover, 2009; Zhu et al., 2015). Signals related to both quasiperiodic respiration (Power et al., 2018) and deep breaths (Kastrup et al., 2001; Power et al., 2017) are apparent within global signal timecourses. Because aCompCor models the principal components among the timeseries of "noise regions" that are unlikely to be driven primarily by neural signal, including such components in nuisance regression is also thought to remove influence of physiological artifacts from respiration and cardiac signal (Behzadi et al., 2007). Thus, even though neither global signal regression (GSR) nor aCompCor isolate respiratory-related signal in a theory-based or model-based way, such nuisance regression strategies have been argued to be the most effective methods for correction of physiological artifacts (Poskanzer et al., 2021; Power et al., 2019; Xifra-Porxas et al., 2021).

On the other hand, global signal regression is a highly debated strategy, and there are potential drawbacks to both global signal regression and component-based approaches. Work in both humans and rodents suggests that the global signal timeseries in part contains meaningful signal arising from neuronal activity (Belloy et al., 2018; Fox et al., 2009; Glasser et al., 2018; Murphy et al., 2009; Thompson et al., 2013; Yousefi et al., 2018). In addition, while global signal regression mitigates associations between head motion and functional connectivity in resting-state fMRI data, it simultaneously introduces distance-dependent motion artifacts such that scan-level head motion (mean FD) is positively correlated with short-distance edges (correlations between two nodes) and negatively correlated with long-distance edges (Ciric et al., 2017b; Power et al., 2014). Some analyses of fMRI data are also reliant on the global signal, and fundamentally incompatible with GSR (Wong et al., 2013). Further, nuisance regression on global signal and noise ROI components can reduce the number of temporal degrees of freedom within the resulting data, particularly when many noise ROI-based components are included (Parkes et al., 2018; Yan et al., 2013).

Censoring high-motion frames, or "scrubbing", has also been proposed as a method to mitigate head motion artifacts in BOLD data (Power et al., 2020; Siegel et al., 2014). Censoring volumes from BOLD data above a motion-based threshold (often framewise displacement >= 0.2mm for resting-state fMRI) has been shown to reduce differences in functional connectivity estimates between high-motion and low motion scans (Power et al., 2013), and mitigate distance-dependent relationships between head motion and functional connectivity (Ciric et al., 2017b). While removal of contaminated frames has demonstrated such positive impacts, respiratory-induced pseudomotion may result in removal of frames with little true head motion (Fair et al., 2020; Gratton et al., 2020; Kaplan et al., 2022), thus unnecessarily excluding data and lowering

statistical power. As such, techniques for mitigating breathing induced psuedomotion are important for removing contaminating influences of head motion without sacrificing uncorrupted volumes.

Encouragingly, recent work has demonstrated that respiratory signals can be estimated accurately from BOLD data alone (Hocke & Frederick, 2021; Salas et al., 2021). In particular, separate stack processing can be used with multiband fMRI datasets to estimate head realignment parameters separately for each 3D set of simultaneously acquired slices (Hocke & Frederick, 2021). A high-resolution set of head realignment parameters can then be constructed by interleaving each separately-estimated timeseries based on the order of acquisition, effectively multiplying the temporal resolution by the number of unique slice times (RF pulses) within each TR. For example, for a dataset with TR=1.4 (~.071Hz) and 16 unique slice acquisition times, a high-resolution timecourse can be estimated at ~11.4Hz. Such high-resolution timecourses are particularly useful for estimating respiratory influence on motion estimates, as they can more accurately capture higher-frequency respiratory signals that are aliased under traditional motion estimation. Respiratory belt traces are more strongly correlated with high-resolution motion estimates (particularly in the phase-encoding direction) than with traditional motion estimation (Hocke & Frederick, 2021).

If high-resolution motion time courses are strongly associated with respiratory traces, then model-based physiological correction strategies (i.e. RETROICOR, RV, RVT) may be possible using these "predicted" respiratory traces without requiring respiratory belt data. While previous research had used BOLD data for accurate reconstruction of low-frequency fluctuations in respiratory volume (RV, Salas et al., 2021), this work is limited by the temporal resolution of the data. Thus, using high-resolution predicted respiratory traces for model-based correction may

allow for better mitigation of higher-frequency artifacts in functional connectivity data. If such correction strategies yield results similar to those when using respiratory belt data, this might eliminate the need for acquiring belt data in new studies, as well as providing a means for informed corrections in existing datasets without peripheral physiological recordings.

In the current study, we asked whether using predicted traces can allow for effective model-based physiological noise correction by comparing an array of different strategies through forked preprocessing pipelines (Bridgeford et al., 2020; Gelman & Loken, 2014). Using resting-state multiband fMRI data from the Nathan Kline Institute Rockland Sample (NKI-RS) and Human Connectome Project (HCP) test-retest sample, we applied model based physiological noise correction (RVT + RETROICOR) using both belt and predicted respiratory traces. In addition to these strategies, we also conducted many preprocessing specifications, varying whether to use a notch filter applied to the head realignment parameters, GSR, aCompCor, and motion-based volume censoring. We evaluate each strategy on benchmarks for data retention (if using censoring), test-retest reliability (I2C2 and discriminability) and residual motion-related artifacts in the preprocessed functional connectivity data.

2.2 Methods

Overview: Implementing methods developed by Hocke & Frederick (2021), we estimated predicted respiratory traces using both the HCP test-retest (Van Essen et al., 2012) and NKI-RS (Tobe et al., 2021) resting-state BOLD data. We next used both belt and predicted traces for model-based respiratory corrections (RVT + RETROICOR), respectively, and compared model-based with frequency-based approaches (notch filtering) for mitigation of pseudomotion artifacts. Then, we preprocessed the data under several forking pipelines, varying whether to use

a notch filter on the head realignment parameters (HCP only), GSR, aCompCor, and motion-based censoring (NKI-RS only). We examined all resultant functional connectomes after preprocessing on benchmarks of test-retest reliability and residual motion-related artifacts.

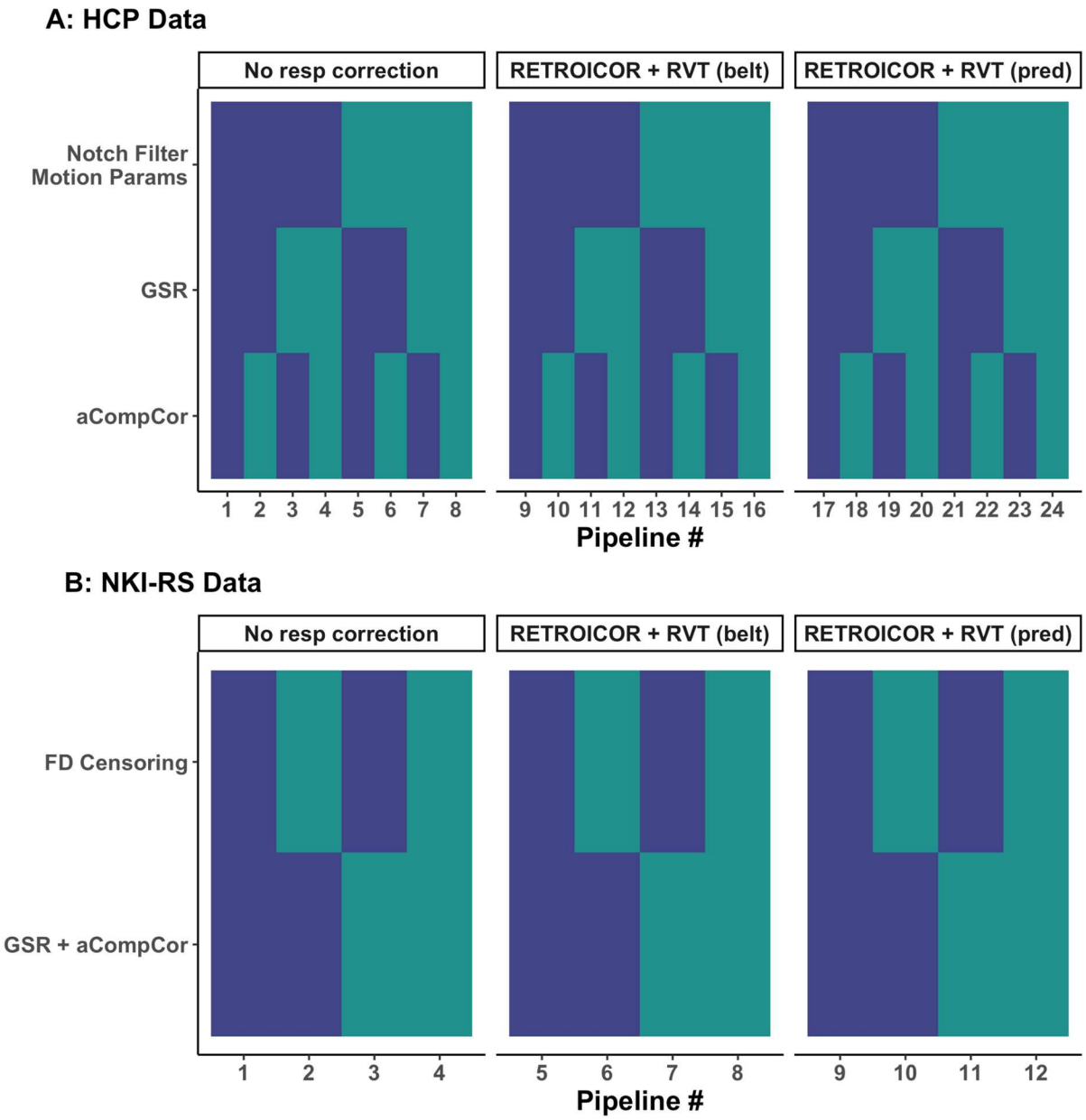


Figure 2.2: Pipeline forks for the HCP (**A**) and NKI-RS (**B**) datasets

Human Connectome Project Data: We used unprocessed BOLD data from the Human Connectome Project (HCP; (Van Essen et al., 2012) test-retest cohort (N = 45). The HCP test-retest participants were scanned at Washington University (St. Louis, USA) using Siemens 3T Skyra for two sessions (Mean between-session interval = 4.7 months, range = [1,11]), each of which spanned two days. For each session, participants completed two 14.4 minute runs each day, 1 with a left-to-right (LR) phase-encoding direction and 1 with a right-to-left (RL) phase-encoding direction. Apart from the flipped phase-encoding direction, all BOLD runs were acquired using the same EPI sequence parameters (TR=0.72s, TE=33.1ms, flip angle=52°, FOV=208x180mm, resolution=2x2x2mm, multiband factor=8) and were 1200 volumes (14.4 minutes) each. During scan acquisition, participants were instructed to keep their eyes open, and a crosshair was shown on the screen. We also used T1w scans (TR=2.4s, TE = 2.14ms, TI = 1s, axial orientation, voxel size 0.7x0.7x0.7mm, flip angle=8°, FOV = 224 × 224 mm) acquired once per session for registration of functional images (Glasser et al., 2013). Within the HCP test-retest cohort, we analyzed data from a total of 36 participants with complete BOLD and respiratory data for all 8 runs across both sessions.

Nathan Kline Institute Rockland Sample Data: This study used multiband resting state fMRI data from the NKI Rockland Sample (NKI-RS) cohort retest session (N = 97, 58F/39M, ages 6-20). We selected a random subsample of participants with complete data from the retest session stratified to match the distribution of mean framewise displacement values within the entire cohort.

NKI-RS participants were scanned at both TR=645ms (TE = 30ms, transversal orientation, voxel size 3.0×3.0×3.0 mm, flip angle=60°, FOV = 222 × 222 mm, 10 interleaved slice acquisition times) and TR=1400ms (TE = 30ms, transversal orientation, voxel size

2.0×2.0×2.0 mm, flip angle=65°, FOV = 224 × 224 mm, 16 interleaved slice acquisition times), both sequences with a multiband acceleration factor of 4. Data were acquired with an anterior-posterior (AP) phase-encoding direction for both sequences. Resting-state scans for both sequences were 10 minutes in duration.

Respiratory belt acquisition: Participants in both the in NKI-RS and HCP samples wore pneumatic respiratory belts around the abdomen during fMRI acquisition, which measured changes in belt tension over time as a proxy for respiration. The elastic abdominal belt was connected to a bellows, such that pressure changes in the bellows were transmitted via a rubber tube for recording. Within the HCP data, belt measurements were taken at a sampling rate of 400 Hz. Belt measurements for the NKI-RS data were taken at a sampling rate of 62.5 Hz. Cardiac measurements were also recorded for both datasets, but not used in the current analyses.

Predicted respiratory traces using high-resolution head motion estimation: We constructed predicted respiratory traces for all scans using methods developed by Hocke & Frederick (2021) and reimplemented using python (Code can be found at https://github.com/pab2163/estimate_respiratory_traces). First, we grouped all simultaneously acquired slices into "stacks," such that the number of slices within each stack was equal to the multiband acceleration factor, and the number of stacks was equal to the number of unique slice acquisition times. We then multiplied the thickness of each slice (in the inferior-superior direction) by the number of unique slice acquisition times. Next, we calculated head realignment parameters separately for each stack using 3DVolreg for rigid-body transformation (Cox, 1996; Teruel et al., 2018). Then, we concatenated head realignment parameters across stacks following the ordering of slice acquisition times to construct high-resolution estimates of head motion, and

applied a notch filter at the original BOLD sampling frequency (as well as 2nd and 3rd harmonics) to remove effects of sampling rate.

Validation of predicted respiratory traces*:* Because breathing-induced signals are particularly apparent within motion estimates in the phase encoding direction, we used these high resolution timeseries in the phase-encoding direction as "predicted" respiratory traces.To validate these BOLD-derived respiratory traces as done by Hocke & Frederick (2021), we computed lagged cross-correlations between the belt trace and high-resolution head motion estimates in each of the three translation axis (left-right, inferior-superior, anterior-posterior). Before computing cross-correlations, we downsampled the belt traces to the sampling frequency of the high-resolution head motion estimates, and filtered both timeseries using a 3rd order bandpass filter of 0.2-0.5 Hz. We then computed all lagged product-moment cross-correlations within a window of +/- 15s (by repeatedly shifting the timeseries relative to one another by 1 sample and zero-padding the corresponding samples in the shifted timeseries) to find the maximum correlation and corresponding lag time for each timeseries with the belt data. In addition to examining such lagged correlations, we also examined the product moment correlations between these timeseries with no lag. Finally, we upsampled the head realignment parameters in the original temporal resolution to the sampling frequency of the high-resolution head motion estimates, and computed similar cross-correlations with the belt traces.

For each dataset, we also examined correspondence between belt and predicted traces in the frequency domain by applying a Savitzky-Golay smoothing filter to the power spectra of each respective trace (https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.savgol_filter.html), then correlating the peak frequencies of these traces across scan runs.

RVT + RETROICOR Physiological Noise Correction: We used both the predicted and belt respiratory traces, respectively, to perform model-based physiological noise correction on the BOLD data. We input traces into RetroTS.py for calculation of both RETROICOR and RVT regressors. RETROICOR regressors are created by estimating the phase within a respiratory cycle (i.e the temporal position from beginning to end of a breath) separately for each slice (Glover et al., 2000). Then, the sine and cosine of the phase, as well as 2 times the phase, are calculated for each slice within each volume. Thus, the 4 RETROICOR regressors for each slice allow for correction of physiological signal at higher frequencies than the BOLD data itself, and can take into account changes in respiration rate over the course of a run (Birn et al., 2008). In contrast, RVT regressors represent lower-frequency fluctuations in breathing rate and envelope. To calculate changes in respiratory volume per time, the breath amplitude is divided by the breath period (interpolated to each volume) for each timepoint (Birn et al., 2006). To account for delayed effects of respiration on BOLD signal, RVT regressors were created for 5 shifts of this timeseries ranging from 0-20 seconds in 4-second intervals. RVT regressors did not vary by slice.

RVT and RETROICOR regressors were constructed with the RetroTS.py program using belt and predicted respiratory traces, respectively. Though RETROICOR can also estimate slice-based regressors to account for cardiac artifacts, we opted not to include these in the current study in order to focus on respiratory-related corrections (Glover et al., 2000). Using afni_proc.py (https://afni.nimh.nih.gov/pub/dist/doc/program_help/afni_proc.py.html), we detrended the RETROICOR and RVT regressors, then regressed them on the BOLD data using 3dREMLfit on a slice-wise basis. Polynomial ‘baseline’ regressors (linear and quadratic,

orthogonalized with RVT and RETROICOR regressors) were included in the regression model. Finally, the residuals from the slice-wise regression model were added to the polynomial baseline to form the corrected BOLD dataset used for further analysis.

For a smaller subset (N=5) of NKI-RS participants, we also conducted the same RVT + RETROICOR procedure using predicted respiratory traces shifted to maximize lagged correlations with the belt timeseries. We then compared head motion power spectra and estimated framewise displacement for these participants for the lagged versus non-lagged predicted respiratory traces, as well as with those when the data were not corrected or corrected with RVT + RETROICOR using the belt traces (see Appendix B Figs. 1-2).

Filtering head realignment parameters: We conducted a rigid-body motion correction of each BOLD run using 3dvolreg (Cox, 1996) to estimate six head realignment parameters (left-right (dL, or 'x'), anterior-posterior (dP, or 'y'), inferior-superior (dS, or 'z'), pitch, roll, and yaw). We then followed methods proposed by Fair et al. (2020) to remove respiratory-induced signal from head realignment estimates by applying a band-stop ('notch') filter to the head realignment parameters for each run. Notch filters can reduce signal around a center frequency within a certain bandwidth of frequencies while leaving signal at other frequencies unaffected. Here, we chose the center frequencies for each dataset as the mean peak predicted respiratory frequency (.31Hz for HCP, .37Hz for NKI-645). Based on Fair et al (2020), we used a bandwidth of .12Hz. Thus, this notch filter was "static", or identical for all scans within each dataset (although scan-specific notch filtering is possible). Because the Nyquist frequency for the NKI-1400 data was ~0.357Hz, we did not apply a notch filter to data for this sequence.

While we only use the static notch filter for full preprocessing of the data, we also filtered the head realignment parameters using a scan-specific notch filter, where the center frequency

was set at the peak predicted respiratory frequency for each scan and the bandwidth was kept at .12 Hz. Notably, this filter was both scan-specific (individualized), and constructed without requiring respiratory belt data. We examined the efficacy of this filter at reducing respiratory artifacts in estimated head motion (see Appendix B Fig. 3).

HCP Preprocessing Specifications: We conducted all preprocessing using C-PAC Version 1.8 (Craddock et al., 2013). We constructed 8 pipeline specifications based on the fMRIPREP-harmonized pipeline (https://github.com/FCP-INDI/C-PAC/blob/develop_v1.8_convergence/CPAC/resources/configs/pipeline_config_fmriprep-options.yml) defaults, varying whether or not to apply a static notch filter to the head realignment parameters before nuisance regression, and whether to include GSR or aCompCor in the nuisance regression (see Table 2.1). We ran all 8 pipelines on raw BOLD runs, BOLD runs passed through belt RVT + RETROICOR, and BOLD runs passed through predicted RVT + RETROICOR such that a common anatomical to MNI registration was used for all 24 total specifications for each given run (see Fig. 2.2).

For functional preprocessing, BOLD data were first motion corrected using 3dvolreg (two passes). At this point, the notch filter was applied to the head realignment parameters generated by 3dvolreg in forks including the step. Next, nuisance regression was applied, including 24 head motion regressors (6 head realignment parameters + squared + delayed + squared delayed; Friston et al., 1996) and linear and quadratic detrending, as well as global signal or aCompCor regressors corresponding with each fork (see Fig. 2.1). Bandpass filtering (.01-.1Hz) was applied to BOLD data after nuisance regression. Functional to anatomical registration matrices were calculated using FSL's boundary-based registration, and anatomical registrations to a MNI template were calculated using ANTS. After all nuisance regression and

filtering, BOLD data were warped to the MNI template and interpolated to 3.4 mm isotropic voxels. No voxel-wise despiking, scrubbing, or spike regression were used in functional preprocessing.

NKI-RS Preprocessing Specifications: Preprocessing for the NKI-RS was largely identical to the above pipelines described for the HCP data but with several exceptions. Within these data we did not include forks with a notch filter applied to the head realignment parameters (although we did estimate impacts of such notch filtering on head motion, see Fig. 2.4), and only included forks for neither GSR nor aCompCor or both within the nuisance regression. Within the NKI-RS data we included an additional forking decision point for whether or not to censor high-motion volumes at the nuisance regression stage. Pipelines with censoring removed volumes with Jenkinson framewise displacement > .2mm (Power et al., 2013, 2014). Thus, between forks for model-based respiratory correction and for nuisance regression specification options, we ran a total of 12 preprocessing pipelines within the NK-RS Data (see Table 2.2).

Decision Point	**# forks**	**Fork Specifications**
Model-based respiratory correction	3	(1) No correction, (2) RVT + RETROICOR (belt), (3) RVT + RETROICOR (predicted)
Notch Filter	2	(1) No filtering of head realignment parameters, (2) notch filter (center = .31Hz, width = .12Hz, order =4) applied to realignment parameters before nuisance regression
Global signal regression	2	(1) No global signal regression, (2) mean signal intensity across all gray matter voxels included as regressor in nuisance regression
aCompCor	2	(1) No aCompCor, (2) top 5 PCs across white matter and cerebrospinal fluid included as regresses in nuisance regression

Table 2.1: Decision forks for the HCP data. All combinations of specifications were run, resulting in 24 (3x2x2x2) pipelines.

Decision Point	# forks	Fork Specifications
Model-based respiratory correction	3	(1) No correction, (2) RVT + RETROICOR (belt), (3) RVT + RETROICOR (predicted)
Global signal regression + aCompCor	2	(1) No global signal regression or aCompcor, (2) mean signal intensity across all gray matter voxels and top 5 PCs across white matter and cerebrospinal fluid included as regresses in nuisance regression
Censoring	2	(1) No censoring, (2) BOLD volumes with FD >= 0.2 censored to remove their influence on functional connectivity estimates

Table 2.2: Decision forks for the NKI-RS data. All combinations of specifications were run, resulting in 12 (3x2x2) pipelines.

Functional connectivity matrices: For all pipelines for each scan run, we extracted the mean timeseries for each of the Schaefer-200 Atlas parcels (Schaefer et al., 2018), then computed product-moment correlations between all pairs of nodes (parcels). For group-level analysis we extracted the upper triangle of all functional connectivity matrices, not including the diagonal.

Power Spectra of Head Realignment Parameters: To visualize impacts of breathing on head motion estimates, we created heatmaps of power spectra of each of the six head realignment parameters across scan runs. Using methods similar to those of Fair et al (2020), we computed the power spectra of each respective parameter using scipy.fftpack.fft (https://docs.scipy.org/doc/scipy/reference/generated/scipy.fftpack.fft.html) for a discrete Fourier transform. We then ranked scan runs from least to greatest median framewise displacement, then plotted heatmaps of relative log power (from 0 up to the Nyquist folding frequency) for all scan runs for each of the six head realignment parameters. We created such heatmaps for five different sets of realignment parameters: (1) raw BOLD data (no correction), (2) after static

notch filtering of the head realignment parameters, (3) after belt RVT+RETROICOR, (4) after predicted RVT+RETROICOR, and (5) after scan-specific notch filtering (see Appendix B Fig. 3). In particular, we examined the heatmaps for the power spectra in the phase-encoding direction (dL for HCP, dP for NKI-RS) to see which pipelines best attenuated signal most likely due to respiratory pseudomotion between about .2-.4 Hz.

Residual head motion artifacts in functional connectivity: We compared the impacts of several correction strategies on estimated head motion (mean framewise displacement), as well as impacts on data retention when applying exclusion criteria based on head motion (i.e. censoring). Specifically we calculated framewise displacement for several strategies: (1) from head realignment parameters based on the raw BOLD data, (2) from head realignment parameters based on the raw BOLD data after applying a static notch filter (except for the NKI-1400 data due to the slower sampling rate), (3) from the head realignment parameters based on BOLD data after RETROICOR + RVT had been applied using the belt respiratory trace ("belt RVT + RETROICOR"), and (4) from the head realignment parameters based on BOLD data after RETROICOR + RVT had been applied using the predicted respiratory trace ("predicted RVT + RETROICOR"). Using these head realignment parameters, we then calculated framewise displacement using both the Power (Power et al., 2012) and Jenkinson (Jenkinson et al., 2002) methods. Power framewise displacement estimates were calculated as the sum of the absolute values of differences from the previous volume across each of the six parameters (assuming a head radius of 50mm for rotation parameters). We calculated Jenkinson framewise displacement using transformation matrices using the CALCULATE_FD_J function from C-PAC (Craddock et al., 2013).

We first compared mean framewise displacement estimates for each sequence (NKI-645, NKI-1400, and HCP) between all four strategies. Secondly, we compared the percentage of volumes with framewise displacement over a threshold FD=0.2mm to measure how much data was censored based on this widely-used criteria (Power et al., 2013; Yan et al., 2013). Third, we compared the percentage of frames under FD=0.2mm in the uncorrected BOLD data that caused FD to be estimated >=0.2mm with each correction strategy. We refer to these frames as "lost," under the logic that applying a given correction strategy combined with censoring at this threshold will result in exclusion of these frames estimated to be low-motion in the uncorrected BOLD data. For each of these three metrics (mean framewise displacement, percentage of volumes censored, percentage of volumes lost), we estimated average differences between pipelines via bootstrapping (resampling participants with replacement, 10000 iterations). For each bootstrap iteration, we compared differences between each pair of pipelines for each individual run, then calculated the mean difference across all runs (see Appendix B Fig. 4 for all comparisons).

Test-retest reliability & discriminability: To quantify how different pipelines impacted the relative similarity of functional connectome measurements within the same participant compared to other participants, we computed image intraclasss correlation coefficients (I2C2) for test-retest reliability. I2C2 is a multivariate generalization of the traditional intraclass correlation coefficient (ICC) for a global Image reliability measure (Shou et al., 2013). Like traditional ICC, I2C2 assumes that observations are drawn from a (multivariate) Gaussian distribution, and is calculated as the proportion of the total variance that is made up of between-participant variance (or more specifically, the trace of the between-participant covariance matrix divided by the trace of the total covariance matrix). We computed all I2C2 calculations using the neuroconductor R

package (Muschelli et al., 2019), and extracted within-participant and between-participant variance estimates as well as I2C2 metric.

Because the I2C2 metric is somewhat limited by Gaussian assumptions and highly sensitive to outliers (Vaz et al., 2013), we also computed multivariate discriminability as a reproducibility metric using the hyppo python package (Panda et al., 2019). Discriminability quantifies the proportion of the time in which multiple measurements of the same item are more similar to one another than they are to other items (Bridgeford et al., 2020). In this context, "items" were represented by participants, and similarity was operationalized through Euclidean distances between pairs of connectivity matrices. Thus, discriminability here represented the proportion of times that different connectivity matrices within the same participant were more similar to one another than with those of other participants (with discriminability=1 representing perfectly discriminable data). Discriminability also has advantages over using a fingerprinting index for reliability (Finn et al., 2015; Milham et al., 2021), because it calculates the proportion of times where measurements from the same participant are more similar to one another then measurements from different participants, rather than using an "all or nothing" approach (as fingerprinting does) to quantify whether the *most* similar measurement is from the same participant or not. We calculated reliability in several contexts.

Reliability between sessions (HCP only): We computed I2C2 and discriminability across functional connectivity matrices for runs (14.4 min each) of the same phase-encoding direction collected during the "test" versus "retest" scanning sessions on separate days. We averaged functional connectivity matrices from pairs of scan runs collected during the same session for each participant with the same phase encoding direction to calculate I2C2 and discriminability for connectomes constructed from 28.8 minutes of data each.

Reliability between sequences (NKI only): To understand which pipelines best harmonized functional connectivity measurements between the TR=1400ms and TR=645 scan sentences, we computed I2C2 and discriminability for runs collected with each sequence for the same participants during the same session.

Inter-pipeline agreement (Reliability between pipelines) (HCP and NKI)**:** To understand which preprocessing choices contributed most to differences in functional connectivity estimates, we computed I2C2 for functional connectivity matrices generated from the same underlying data but with varying preprocessing pipelines. Here, pipelines were treated as "raters", and we calculated I2C2 and discriminability between all pairs of pipelines.

For reliability between sessions and sequences, we quantified the sampling variability in I2C2 and discriminability through bootstrapping participants. For each of 1000 bootstrap iterations, we drew a random sample (without replacement) of two-thirds of the participants in each dataset for which to calculate each metric.

Head motion correlations with functional connectivity: To examine relationships between head motion and functional connectivity after preprocessing, we computed framewise displacement-functional connectivity (FD-FC) correlations for each pipeline. We calculated rank-order correlations between each edge weight (representing functional connectivity between two nodes) and Jenkinson mean FD each scan (Power et al., 2012; Satterthwaite et al., 2012). Within the HCP data, we computed FD-FC correlations across all scan runs for each respective pipeline. Within the NKI-RS data, we computed such correlations separately for each of the two scan acquisition sequences. For each pipeline, we summarized distributions of signed FD-FC correlations as well as distributions of correlation magnitudes (absolute value; Ciric et al.,

2017b). For statistical comparisons of FD-FC relationships across pipelines, we conducted 1000 iterations of bootstrap resampling to calculate distributions of median FD-FC across all edges.

Distance-Dependence of head motion correlations with functional connectivity: Head motion during scanning can influence functional connectivity estimates in a distance-dependent way, such that connectivity for nearby nodes is increased and distant nodes decreased (Power et al., 2012, 2014). Thus, we sought to understand how head motion artifacts varied as a function of distance between nodes under each pipeline. To examine the impacts of preprocessing choices on these distance-dependent artifacts, we first calculated Euclidean distances between the center of mass for all pairs of nodes in the Schaefer 200 atlas. We then calculated rank-order correlations between the distance separating each pair of nodes (edge distance) and FC-FD correlation values for each edge. We conducted analyses for all pipelines across all scans within the HCP data, and separately for each of the two scan acquisition sequences within the NKI-RS data. For statistical comparisons of distance-dependent FD-FC relationships across pipelines, we conducted 1000 iterations of bootstrap resampling to calculate distributions of the correlation between distance and FD-FC across all edges.

2.3 Results

Validation of predicted respiratory traces: We asked whether high-resolution motion estimates in the phase-encoding direction were effective as "predicted" respiratory traces, as previously demonstrated by Hocke & Frederick (2021), by examining their similarity with corresponding respiratory belt traces. Maximum lagged cross-correlations between belt and predicted (high-resolution motion estimates in the phase-encoding direction) traces within the 0.2-0.5 Hz range (Fig. 2.3A-B) had medians of 0.63, 0.69, 0.91 in the NKI-1400, NKI-645, and

HCP data, respectively. Although such cross-correlations in the NKI-RS data were not as strong as those previously reported in a different subset of NKI-RS scans by Hocke & Frederick (2021), such relationships may have been weakened by the inclusion of runs with higher head motion and lower quality respiratory belt data. Indeed, within the NKI-RS data, correspondence between belt traces and high-resolution motion estimates was stronger for scans with lower head motion estimates and less clipping of the respiratory signal (see Appendix B Fig. 5).

While for most scans, the temporal shift (lag) that maximized the correlation between the belt and predicted respiratory trace was under 2.5s (see Appendix B Fig. 6), we also computed similar correlations without such temporal lags. Correlations between the belt trace and high-resolution head motion in phase-encoding direction without lag were somewhat weaker, though still fairly strong, especially in the HCP data (see Appendix B Fig. 7).

Similarly, we found strong associations between the peak frequencies of belt traces and respective predicted traces using high-resolution motion estimation. For most HCP scans (Fig. 2.3D), peak frequencies were nearly identical between belt and predicted traces, while some NKI scans demonstrated greater discrepancies in peak frequency.

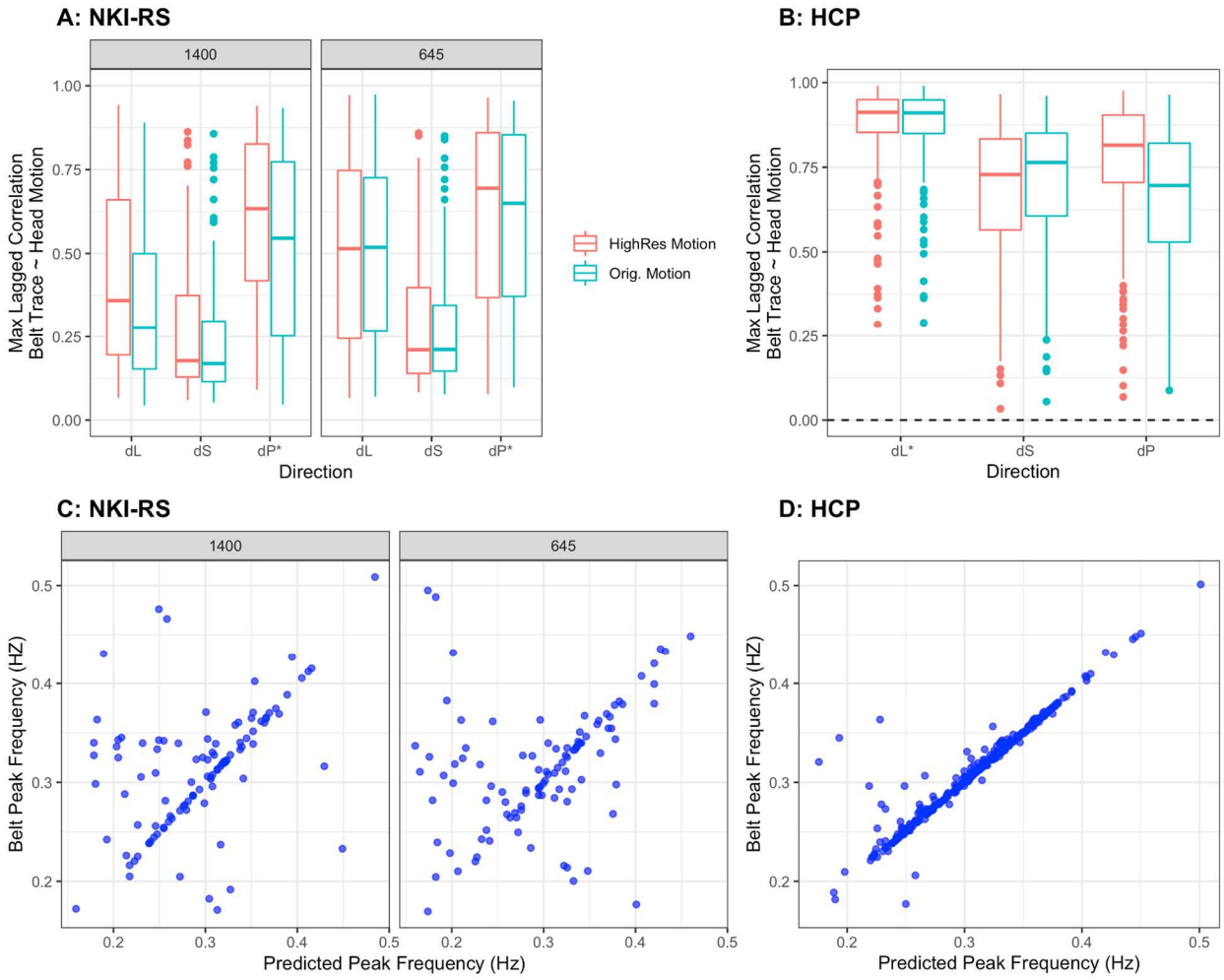


Figure 2.3: Validation of predicted respiratory traces in the NKI-RS and HCP data. **A-B:** Maximum lagged correlations within a +/-15s window between the respiratory belt and head motion estimates in each axis of translation (red = high-resolution motion estimates, blue = original motion estimates). Boxplots of correlations are shown for the NKI data for the TR=1400ms sequence (A, left panel) and TR=645ms sequence (A, right panel) and HCP data (B). High-resolution motion estimates in the phase-encoding direction (dP for NKI-RS, dL for HCP) are treated as the predicted respiratory timeseries. **C-D:** Scatter plots displaying associations between the peak frequency in predicted (x-axis) and belt (y-axis) respiratory traces. Each point represents 1 scan run, and points are shown for the NKI data for the TR=1400ms sequence (C, left panel) and TR=645ms sequence (C, right panel) and HCP data (D)

Impacts of correction strategies on estimated head motion: To assess impacts of respiratory correction strategies on estimated head motion and data retention, we compared framewise displacement estimates for the uncorrected BOLD data to those after applying a notch

filter to the head motion parameters, and RVT + RETROICOR using the belt and predicted traces, respectively. We calculated framewise displacement estimates using both Power (Power et al., 2012) and Jenkinson (Jenkinson et al., 2002) methods, and note that head motion estimates were consistently higher using the Power calculation (Fig. 2.4A & D). Prior work has indicated that Jenkinson calculations are likely more accurate (Yan et al., 2013). Mean framewise displacement estimates using the Power calculation were on average 1.71 times (SD = 0.10) higher than when using the Jenkinson calculation in the NKI data and 1.81 times (SD = 0.10) higher in the HCP data. Apart from Figure 2.4, we focused on Jenkinson framewise displacement estimates, and all reported statistics use Jenkinson calculations unless otherwise noted.

All correction strategies generally reduced mean framewise displacement relative to the uncorrected BOLD data, with the notch filter and predicted RVT + RETROICOR showing the least head motion (Fig. 2.4A & D, see Appendix B Fig. 4 for all statistics). Although head motion estimates were slightly lower using predicted RVT + RETROICOR compared to the notch filter in the HCP data (Mean Difference$_{HCP}$ = -.003mm, 95% CI [-.005, -.001], p = .0003), we did not observe a consistent difference in the NKI-645 data (Mean Difference$_{NKI\text{-}645}$ = 0.001mm , 95% CI [-.003, .007], p = .606). Both predicted RVT + RETROICOR and the notch filter resulted in lower estimated mean framewise displacement compared to belt RVT + RETROICOR. Framewise displacement estimates were higher for the NKI-1400 compared to the NKI-645 data, although this discrepancy was mitigated by normalizing to framewise displacement per minute (see Appendix B Fig. 8).

To estimate impacts of correction strategies on data retention, we calculated the proportion of volumes in each scan over a threshold of Jenkinson framewise displacement >= 0.2mm. Most generally, correction strategies reduced the proportion of volumes that would be

censored using this threshold, although reductions were of greater magnitude in the NKI-1400 where motion estimates were highest. The notch filter (Mean Difference$_{NKI-645}$ = 1.38%, 95% CI [0.74, 2.13], p = .0001; Mean Difference$_{HCP}$ = 5.74%, 95% CI [4.64, 6.93], p = .0001) and predicted RVT + RETROICOR (Mean Difference$_{NKI-645}$ = 0.99%, 95% CI [0.08, 1.9], p = .0349; Mean Difference$_{NKI-1400}$ = 8.99%, 95% CI [6.32, 12.00], p =.0001; Mean Difference$_{HCP}$ = 5.94%, 95% CI [4.73, 7.26], p = .0001) strategies resulted in the greatest reductions in volumes censored relative to no correction, though the proportion of volumes over threshold was not different between these two strategies (Mean Difference$_{NKI-645}$ = 0.39%, 95% CI [0.06, 1.02] p = .109; HCP Mean Difference = -0.20%, 95% CI [-0.50, 0.14], p = .252).

We also examined how often correction strategies caused volumes below the 0.2mm framewise displacement threshold in the uncorrected BOLD data to pass above this threshold after correction (i.e. "lost" TRs). We calculated the proportion of all volumes in each run that were lost under each correction strategy applied. All correction strategies resulted in lost TRs compared to no correction, though using Jenkinson FD, generally only a small percentage were lost. The proportion of lost volumes was higher when using the notch filter compared to the predicted RVT + RETROICOR in both the NKI-645 (Mean Difference$_{NKI-645}$ = -0.66%, 95% CI [-1.11, -0.08], p = .0273) and HCP (Mean Difference$_{HCP}$ = -0.50%, 95% CI [-0.69, -0.335], p = .0001), data, as well as the belt RVT + RETROICOR strategy in the HCP (Mean Difference$_{HCP}$ = -0.36%, 95% CI [-0.55, -0.20], p = .0001), but not NKI-645 data (Mean Difference$_{NKI-645}$ = -0.35%, 95% CI [-1.06, 0.65], p = .403). Visualization of individual head motion timeserieses revealed that the notch filter may tend to "spread" rapid peaks in head motion, such that FD estimates for low-motion frames occurring immediately before or after rapid large movements are inflated (see Appendix B Fig. 9).

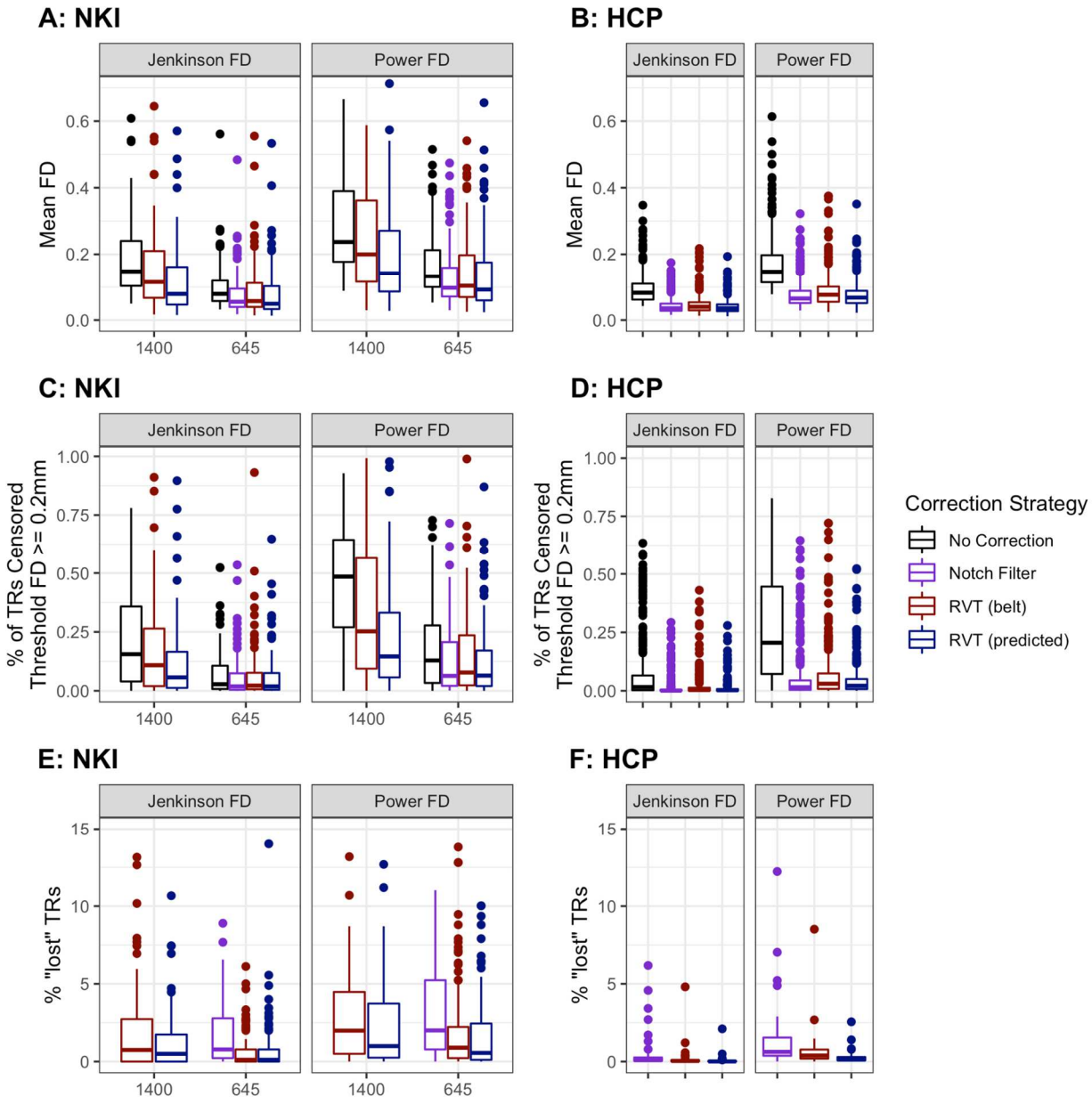


Figure 2.4: Head motion estimates under different correction strategies. **A-B**: Boxplots of mean framewise displacement estimates under both the Jenkinson and Power methods using each correction strategy (black = no correction, purple = notch filter, red = RVT + RETROICOR (belt), blue = RVT + RETROICOR (predicted). **C-D:** Boxplots displaying the proportion of volumes censored under each strategy using a FD >= 0.2mm threshold. **E-F:** Boxplots showing the percentage of volumes "lost" with each correction strategy relative to no correction. Lost volumes are calculated as the percentage of all volumes with FD < 0.2mm without correction, but with FD >= 0.2 with a correction applied. All plots show estimates shown for the NKI-RS (left) and HCP (right) data. Notch filtering was not performed for the NKI 1400 sequence

because much of the filter envelope (center = .37Hz, width = .12Hz) was above the Nyquist frequency (~0.357Hz) of the BOLD sampling rate.

Power spectra of head realignment parameters: To visually inspect impacts of respiration-induced pseudomotion within the head realignment parameters, we constructed heatmaps of the power spectra of each of the 6 head realignment parameters. In the HCP data, while without any correction, strong signatures of both respiratory induced motion (or pseudomotion) were visible in the .2-.4Hz range, particularly in the phase encoding direction (Fig. 2.5A). While a static notch filter decreased some signal in this frequency band (Fig. 2.5B), it also "missed" signal within this range for some participants. RETROICOR + RVT, whether using belt (Fig. 2.5C) or predicted (Fig. 2.5D) respiratory traces, tended to remove such signal over a smoother frequency range. Results were similar for the NKI-RS data, although visualization of respiratory artifacts in the NKI-1400 sequence was somewhat less precise due to the Nyquist frequency (~0.357Hz) falling within the frequency band of such artifacts (see Appendix B Figs. 10-11). We also note a frequency band at ~0.12Hz only in the phase encoding direction, which in prior work has been suggested to represent the influence of slower, deep breaths (Power et al., 2020). However, the fact that this frequency band seems to appear only in datasets collected on Siemens Prisma (or Skya with Prisma-like hardware used in HCP) scanners suggests that this may be a scanner-driven or sequence-driven artifact (see Appendix B Fig. 12; (Kaplan et al., 2022; Power et al., 2019).

Power Spectra of Head Realignment Parameters (HCP Data)

A

No Correction

B

Notch filter
Center = .31Hz
Width = .12Hz

C

RETROICOR + RVT
(belt)

D

RETROICOR + RVT
(predicted)

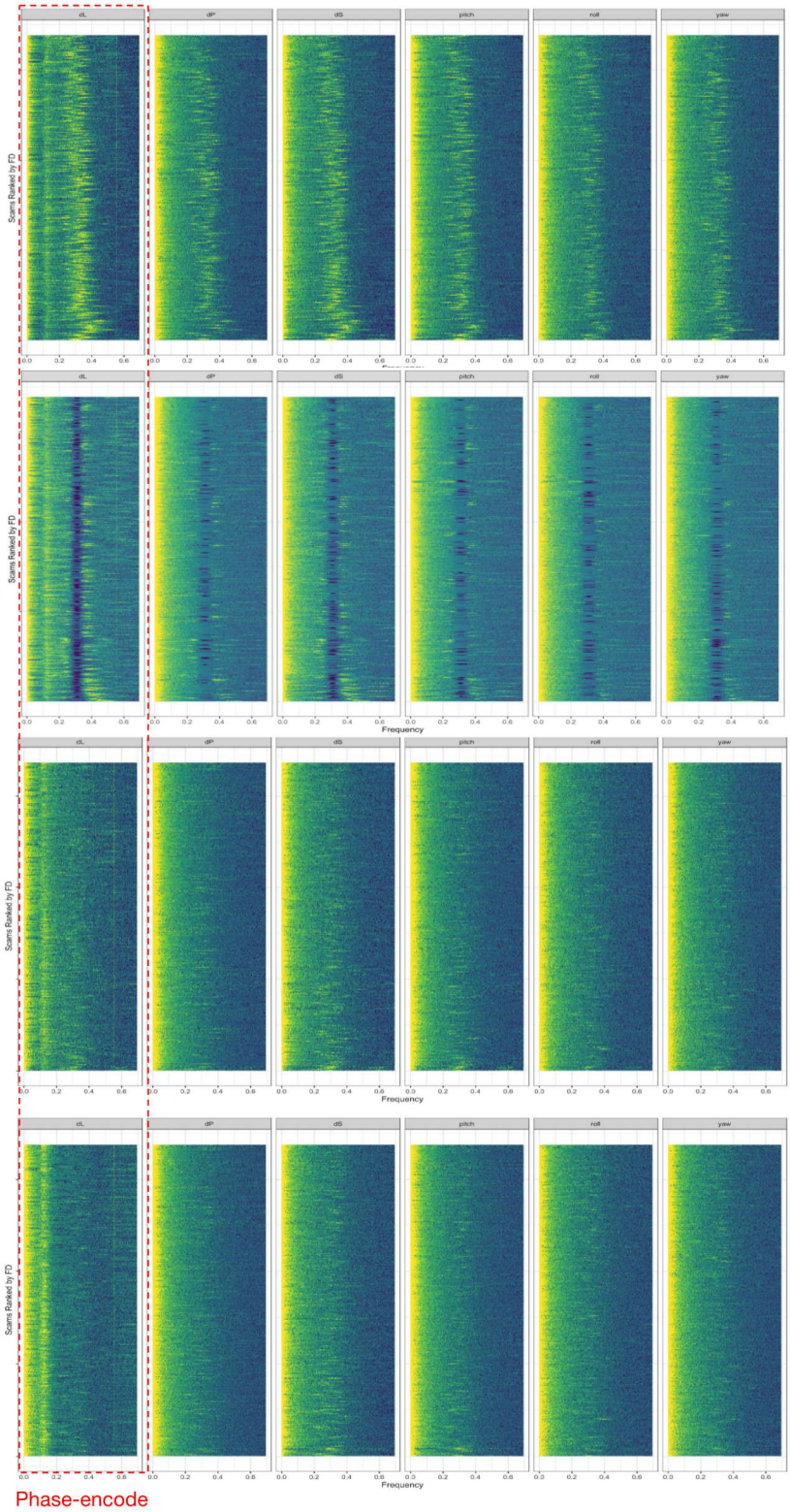


Figure 2.5: Power spectra of head realignment parameters in the HCP dataset with **(A)** no correction, **(B)** a notch filter of the head realignment parameters, **(C)** RETROICOR + RVT with belt trace, and **(D)** RETROICOR + RVT with the predicted respiratory trace. For each plot, each row represents one scan run, with runs ranked from lowest median framewise displacement (top) to highest (bottom). Columns represent the 6 head motion parameters, where pitch, roll, and yaw indicate rotation and dS (inferior-superior), dL (left-right), and dP (anterior-posterior) indicate translation.

Test-Retest Reliability of Functional Connectivity Between Sessions: Using the HCP data, we compared bootstrapped distributions for the test-retest reliability of connectivity matrices across sessions using both the I2C2 and discriminability metrics. We report statistics for scans acquired in the LR phase encode direction in the main text, and RL results were highly similar (see supplement). Neither notch filtering (ΔI2C2 = 0.002, 95% CI [-0.002, 0.005], p =413; $\Delta_{discrim}$ = -0.001, 95% CI [-0.004, 0.002], p = 0.742) nor RVT + RETORICOR using belt (ΔI2C2 = 0.001, 95% CI [-0.009, 0.011]; $\Delta_{discrim}$ = 0.004, 95% CI [-0.005, 0.012], p = 0.413) or predicted (ΔI2C2 = -0.002, 95% CI [-0.011, 0.0069]; $\Delta_{discrim}$ = 0.001, 95% CI [-0.006, 0.008], p = 0.792) respiratory traces had a significant impact on either I2C2 (Fig. 2.6A, right panel) or discriminability (Fig. 2.7A). Within-participant variance (Fig. 2.6A, left and center panels) was lower for pipelines using both belt (Δ_{within} = -11.1, 95% CI [-15.1, -7.07]) and predicted (Δ_{within} = -7.65, 95% CI [-11.5, -3.79]) RVT + RETROICOR, as well as notch filtering (Δ_{within} = -1.31, 95% CI [-2.67, -0.01]), although effects of notch filtering on within-participant variance were smallest in magnitude. However, both belt ($\Delta_{between}$ = -10.7, 95% CI [-17.7 , -2.93]) and predicted ($\Delta_{between}$ = -10.1, 95% CI [-15.8, -3.53]) RVT + RETROICOR decreased between-participant variance.

GSR and aCompCor impacted test-retest reliability much more than did RVT + RETROICOR or notch filtering strategies. The combination of GSR and aCompCor resulted in

the highest I2C2 compared to other pipelines (ΔI2C2 = 0.053, 95% CI [0.012, 0.089]), although GSR without aCompCor resulted in the lowest (ΔI2C2 = -0.066, 95% CI [-0.115, -0.007]) I2C2 estimates relative to others (Fig. 2.6A, right panel). While GSR did not consistently impact I2C2 overall (ΔI2C2 = -0.010, 95% CI [-0.057, 0.037]), this step both reduced within-participant (Δ_{within} = -34.0, 95% CI [-57.7, -9.49]) variance and between-participant variance ($\Delta_{between}$ = -46.4, 95% CI [-75.0, -18.8]; Fig. 2.6A, left panel). aCompCor did not impact I2C2 consistently (ΔI2C2 = 0.0467, 95% CI [-0.005, 0.103]) or between-participant variance ($\Delta_{between}$ = -16.4, 95% CI [-41.1, 12.0]), but did decrease within-participant variance (Δ_{within} = -51.4, 95% CI [-78.8, -17.1]). GSR and aCompcor also reduced the impacts of sampling variability on both I2C2 and discriminability across bootstrap resamples (i.e. lower variance across bootstraps). GSR increased discriminability ($\Delta_{discrim}$ = 0.062, 95% CI [0.024, 0.089], p = 0.001), while aCompCor increased discriminability specifically in pipelines with GSR ($\Delta_{discrim}$ = 0.046, 95% CI [0.001, 0.071], p = 0.029), but did not significantly impact discriminability in pipelines without GSR ($\Delta_{discrim}$ = 0.016, 95% CI [-0.027, 0.058], p = 0.493; Fig. 2.7A). Figures 2.6-7 display results for the HCP scans acquired in the left-right phase-encoding direction, though results were highly similar for scans acquired in the right-left phase encoding direction (see Appendix B Figs. 13-14).

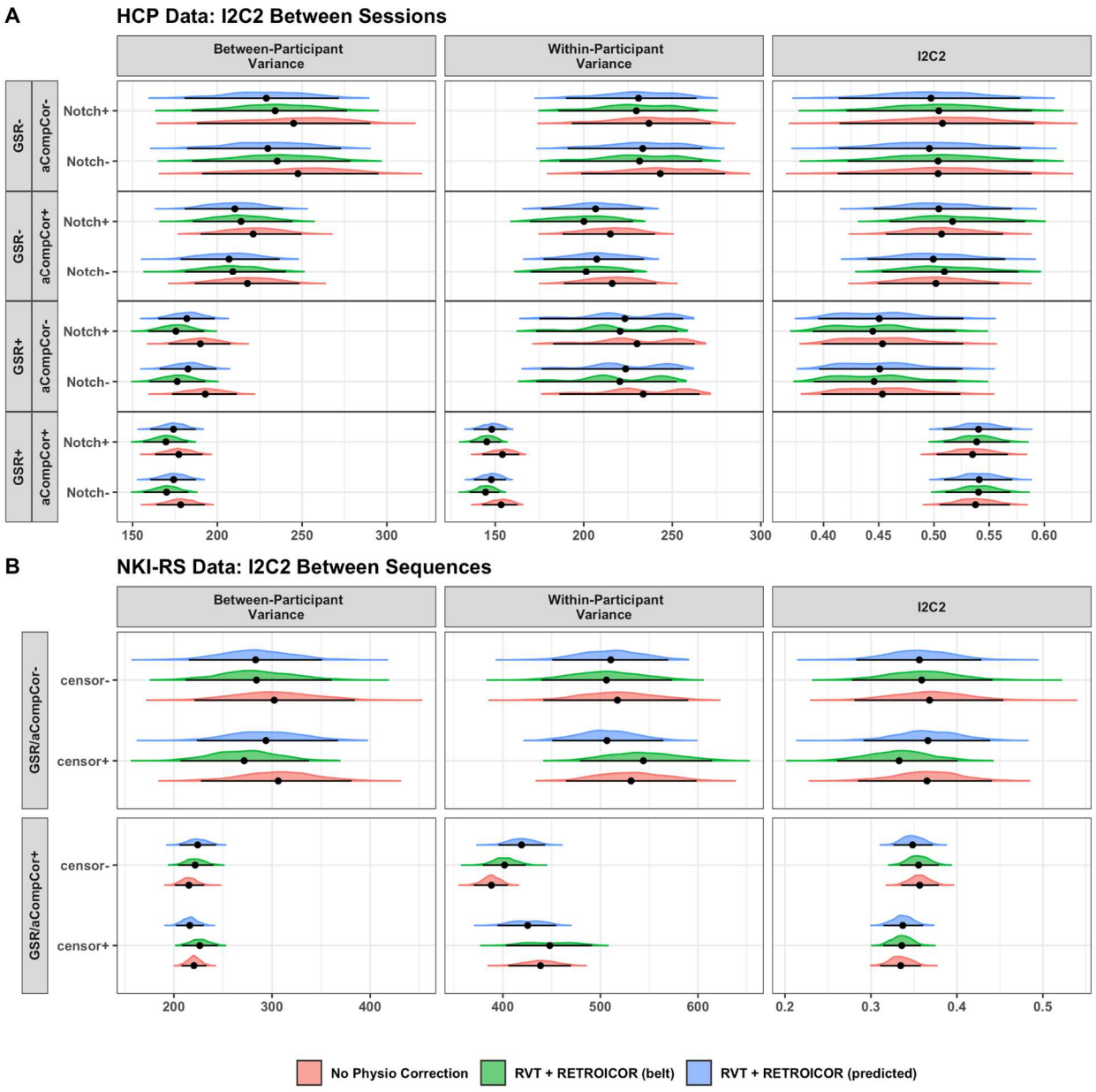


Figure 2.6: A: I2C2 test-retest reliability of functional connectivity matrices between sessions in the HCP data. Data shown here are scans with left-right phase-encoding (see Appendix B Fig. 13 for right-left phase-encoding). Y-axis indicates the chosen forks for each pipeline (GSR, aCompCor, notch filtering head motion parameters). **B.** I2C2 test-retest reliability of functional connectivity matrices between the NKI-645 and NKI-1400 sequences, within the same session. Y-axis indicates chosen forks for each pipeline (GSR/aCompcor together, censoring volumes based on framewise displacement). For all plots, color indicates whether a model-based correction was used (red = no physio correction, green = RVT + RETROICOR (belt), blue = RVT + RETROICOR (predicted). The left two panels show between-participant variances (left) and within-participant variances (center) in arbitrary units, and the right panel shows I2C2 values

(0 = reliability, 1 = perfect reliability). Shaded densities indicate bootstrapped distributions using repeated random resampling of ⅔ of the participants, and points with error bars represent bootstrapped means and 95% confidence intervals. Note: x-axis scales differ between panels A & B in order to more clearly highlight comparisons within each dataset.

Test-Retest Reliability of Functional Connectivity Between Sequences: Using the NKI-RS data, we compared bootstrapped distributions for the test-retest reliability of connectivity matrices acquired during the same session using TR=645ms versus TR=1400ms sequences. As with the HCP data, RVT + RETROICOR using belt (ΔI2C2 = -0.010, 95% CI [-0.024, 0.003]) or predicted (ΔI2C2 = -0.004, 95% CI [-0.021, 0.013]) respiratory traces did not impact estimates of I2C2 (Fig. 2.6B, right panel). Neither belt nor predicted RVT + RETROICOR impacted within-participant variance (Δ_{within_belt} = 6.09 95% CI [-11.7, 24.2], $\Delta_{within\_predicted}$ = -3.44, 95% CI [-23.1, 16.1]) nor between-participant variance ($\Delta_{between\_belt}$ = -10.1 95% CI [-22.6, 3.25], $\Delta_{between\_predicted}$ = -6.69, 95% CI [-20.3, 6.59]. Neither belt ($\Delta_{discrim}$ = -0.008, 95% CI [-0.019, 0.003], p = .173) nor predicted ($\Delta_{discrim}$ = -0.006 , 95% CI [-0.019, 0.007], p = .397) RVT + RETROICOR impacted discriminability (Fig. 2.7B).

As with between-session test-retest reliability, the combination of GSR and aCompCor had the largest effects on reliability between the NKI-645 and NKI-1400 sequences. GSR+aCompCor reduced both between-participant variance ($\Delta_{between}$ = -69.5 , 95% CI [-136.1, -7.05]) and within-participant (between-sequence, Δ_{within}= -99.2, 95% CI [-152.0, -48.7]; Fig. 2.6B, left/center panels). While pipelines using GSR+aCompCor showed the numerically lowest I2C2 estimates (Fig. 2.6B, right panel), GSR+aCompCor did not consistently impact I2C2 (ΔI2C2 = -0.013 95% CI [-0.078, 0.052]). In contrast to I2C2, GSR+aCompCor also increased estimates of discriminability between sequences ($\Delta_{discrim}$ = 0.174 , 95% CI [0.142, 0.206], p = .0001; Fig. 2.7B). Further, pipelines including GSR+aCompCor showed lower variance in

discriminability across bootstrap resamples, indicating a reduced influence of sampling variability on the discriminability under such pipelines. Censoring did not impact between-sequence discriminability ($\Delta_{discrim}$ = -0.006 , 95% CI [-0.0213, 0.009], p = .479).

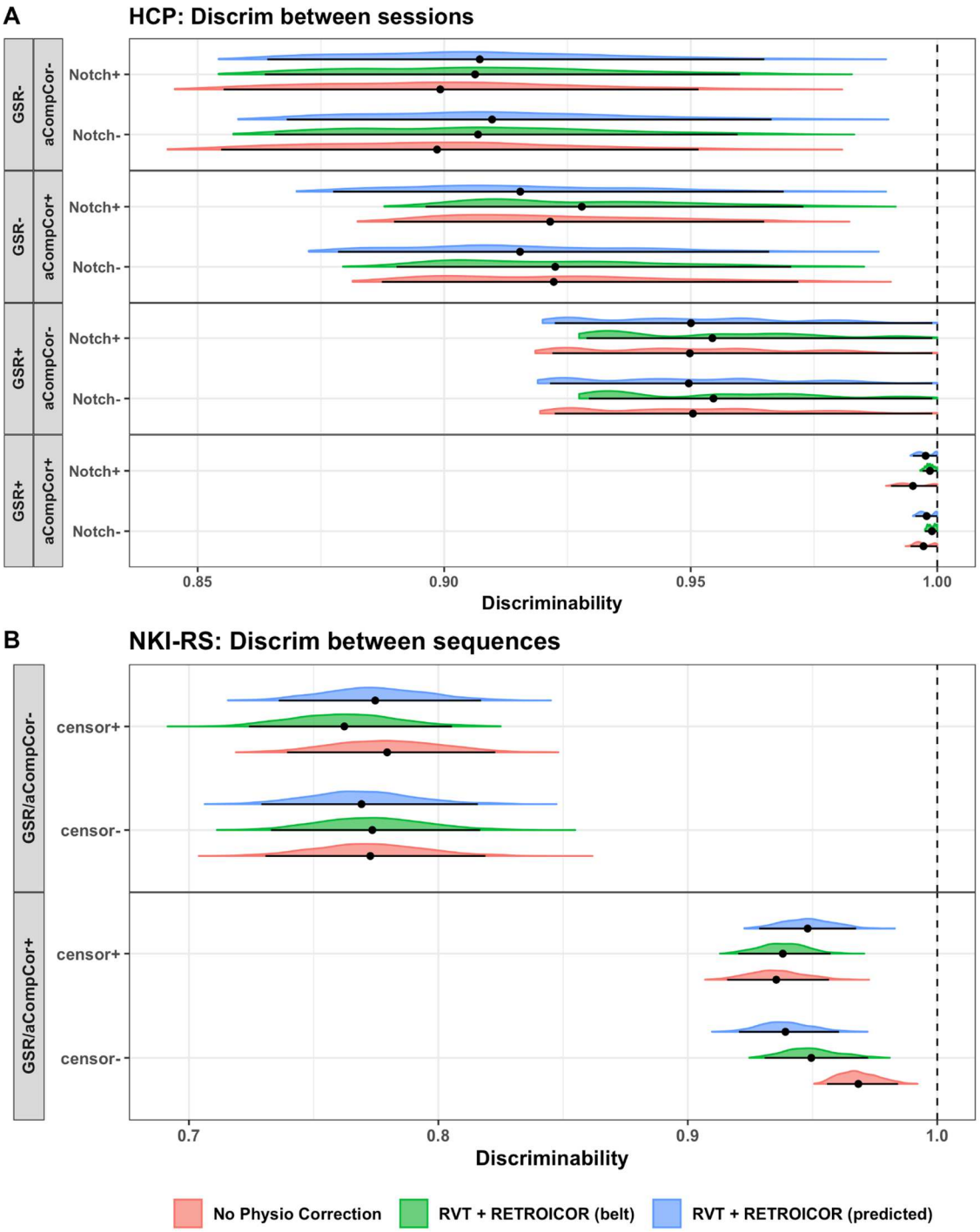


Figure 2.7: **A:** Test-retest discriminability of functional connectivity matrices between sessions in the HCP data. Data shown here are scans with left-right phase-encoding (see Appendix B Fig. 14 for right-left phase-encoding). Y-axis indicates the chosen forks for each pipeline (GSR, aCompCor, notch filtering head motion parameters). **B.** Test-retest discriminability of functional connectivity matrices between the NKI-645 and NKI-1400 sequences, within the same session. Y-axis indicates chosen forks for each pipeline (GSR/aCompcor together, censoring volumes

based on framewise displacement). For all plots, color indicates whether a model-based correction was used (red = no physio correction, green = RVT + RETROICOR (belt), blue = RVT + RETROICOR (predicted). Shaded densities indicate bootstrapped distributions using repeated random resampling of ⅔ of the participants, and points with error bars represent bootstrapped means and 95% confidence intervals. A value of 1 indicates perfect discriminability. Note: x-axis scales differ between panels A & B in order to more clearly highlight comparisons within each dataset.

Inter-pipeline agreement: To quantify how different pipeline choices contributed most to variability in functional connectivity, we computed I2C2 for functional connectivity estimates generated from the same data between all pairs of pipelines. Within the HCP data, the order in which pipeline choices contributed to variability, from most to least, was GSR > aCompCor > RVT + RETROICOR > notch filter (Fig. 2.7B, bottom panel). I2C2 was always lowest between pairs of pipelines that differed (i.e. one pipeline with GSR, one without) on whether GSR was applied (ΔI2C2 = .323, 95% CI [.315, .331]). Varying aCompCor also contributed to variability (ΔI2C2 = .102, 95% CI [.094, .110]). In contrast, varying RVT + RETROICOR only contributed to minimal decreases in I2C2 (ΔI2C2 = .017, 95% CI [.008, .025]), and varying the notch filter step did not consistently decrease I2C2 (ΔI2C2 = .004, 95% CI [-.003, .012]). I2C2 estimates of reliability both between sessions and between pipelines demonstrated similar results.

Within the NKI data, the order in which the pipeline choices contributed to variability, from most to least, was GSR+aCompCor > RVT + RETROICOR > censoring. I2C2 was always lowest between pairs of pipelines that differed on whether GSR+aCompCor was applied (ΔI2C2 =.447, 95% CI [.435, .459]). To a lesser extent, varying RVT + RETROICOR (ΔI2C2 =.057, 95% CI [.044, .069]) and censoring (ΔI2C2 =.043, 95% CI [.031, .055]) also decreased between-pipeline I2C2.

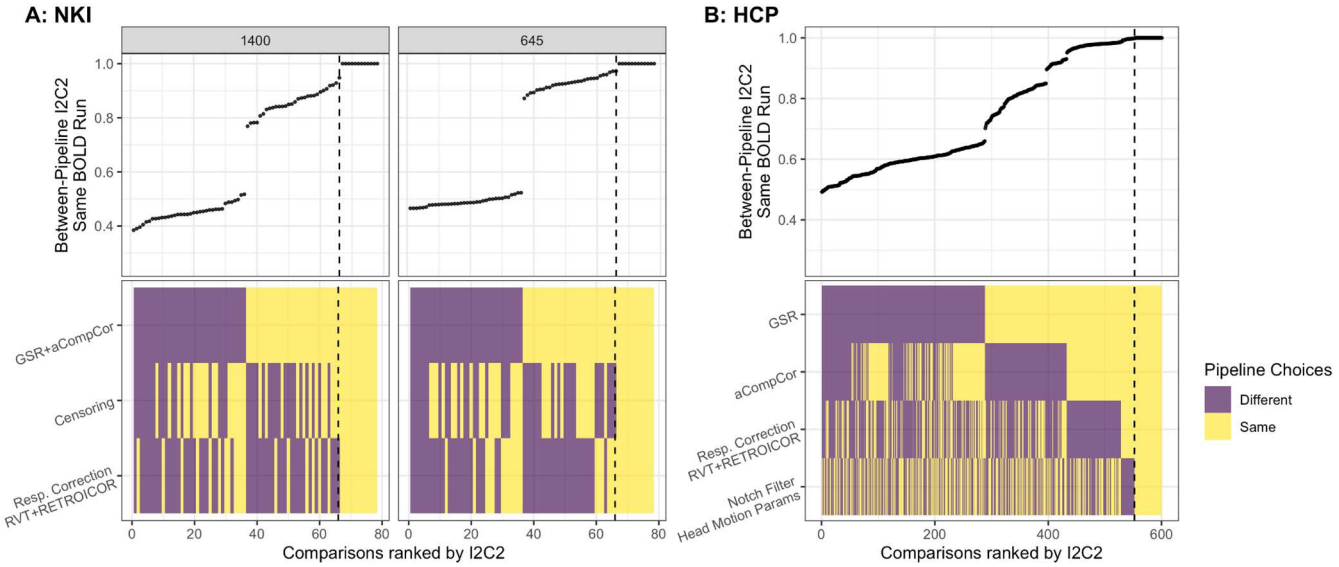


Figure 2.8: Inter-pipeline agreement as measured by I2C2 between pipelines. I2C2 estimates (top) are shown for identical BOLD data processed using different combinations of pipelines. X-axis represents each I2C2 measurement comparing two pipelines, and the heatmaps (bottom) show whether the choice for any given decision point was the same (yellow) or different (purple). Comparisons to the right of the dotted line indicate identical preprocessing pipelines as well as BOLD data (this is why all such I2C2 estimates are 1). **A:** Between-pipeline I2C2 estimates in the NKI data for the 1400 (left) and 645 (right) sequences. **B:** Between-pipeline I2C2 estimates in the HCP data for data in the left-right phase encoding direction.

FD-FC Correlations: We computed correlations across scans between mean framewise displacement (FD) and functional connectivity edge weights between all pairs of nodes to examine relationships between motion and functional connectivity (FD-FC) after preprocessing (Fig. 2.9). Overall, impacts of notch filtering and RVT + RETROICOR on FD-FC correlations were relatively minor. Within the HCP data, notch filtering did not impact FD-FC relationships ($\Delta_{abs(FD\text{-}FC)}$ = 0.003, 95% CI [-0.016, 0.024], p = 0.784). Within the HCP data neither RVT + RETROICOR using belt ($\Delta_{abs(FD\text{-}FC)}$ = 0.001, 95% CI [-0.018, 0.020], p = 0.994) nor predicted ($\Delta_{abs(FD\text{-}FC)}$ = -0.002, 95% CI [-0.020, 0.018], p = 0.820) traces impacted FD-FC relationships. Within the NKI-RS data, neither RVT + RETROICOR using belt (ΔNKI-645$_{abs(FD\text{-}FC)}$ = -0.002,

95% CI [-0.059, 0.055], p = 0.938; ΔNKI-1400$_{abs(FD\text{-}FC)}$ = 0.013, 95% CI [-0.040, 0.072], p = 0.622) nor predicted (ΔNKI-645$_{abs(FD\text{-}FC)}$ = 0.006, 95% CI [-0.051, 0.065], p = 0.874; ΔNKI-1400$_{abs(FD\text{-}FC)}$ = 0.003, 95% CI [-0.054, 0.060], p = 0.918) traces impacted FD-FC relationships most generally. RTV + RETROICOR using both belt (ΔNKI-645$_{signed(FD\text{-}FC)}$ = 0.046, 95% CI [0.021, 0.074], p = 0.001; ΔNKI-1400$_{signed(FD\text{-}FC)}$ = 0.046, 95% CI [0.025, 0.068], p = 0.001) and predicted traces did increase signed FD-FC correlations specifically among pipelines with GSR+aCompCor (Fig. 2.9B, right panel). However, neither RVT + RETROICOR using belt (ΔNKI-645$_{abs(FD\text{-}FC)}$ = 0.015, 95% CI [-0.007, 0.037], p = 0.207; ΔNKI-1400$_{abs(FD\text{-}FC)}$ = 0.016, 95% CI [-0.007, 0.038], p = 0.135) nor predicted (ΔNKI-645$_{abs(FD\text{-}FC)}$ = 0.016, 95% CI [-0.006, 0.038], p = 0.167; ΔNKI-1400$_{abs(FD\text{-}FC)}$ = 0.018, 95% CI [-0.004, 0.041], p = 0.085) traces impacted the magnitude of FD-FC correlations in pipelines with GSR+aCompCor (or across all pipelines; Fig. 2.9B, left panel).

GSR, compared to all other decision points, had the largest impacts on FD-FC correlations. Across all datasets, pipelines including a GSR step showed distributions of signed FD-FC correlations centered closer to 0 (Fig. 2.9, right panel). Within the HCP data, GSR reduced the magnitude of FD-FD correlations ($\Delta_{abs(FD\text{-}FC)}$ = -0.013, 95% CI [-0.024, -0.003], p = 0.007). Within the NKI data, pipelines with GSR+aCompCor also showed reduced magnitude of FD-FC correlations (ΔNKI-645$_{abs(FD\text{-}FC)}$ = -0.089, 95% CI [-0.137, -0.042], p = 0.001; ΔNKI-1400$_{abs(FD\text{-}FC)}$ = -0.147, 95% CI [-0.192, -0.102], p = 0.001). Within the HCP data, aCompCor alone did not influence FD-FC correlations ($\Delta_{abs(FD\text{-}FC)}$ = 0.008, 95% CI [-0.003, 0.019], p = 0.187), although it did reduce signed correlations specifically in pipelines with GSR ($\Delta_{signed(FD\text{-}FC)}$ = -0.008, 95% CI [-0.012, -0.004], p = 0.001). Censoring reduced the magnitude of FD-FC

correlations within the NKI-1400 data ($\Delta_{abs(FD-FC)}$ = -0.076, 95% CI [-0.122, -0.025], p = 0.003), but not within the NKI-645 data ($\Delta_{abs(FD-FC)}$ = -0.006, 95% CI [-0.053, 0.041], p = 0.796).

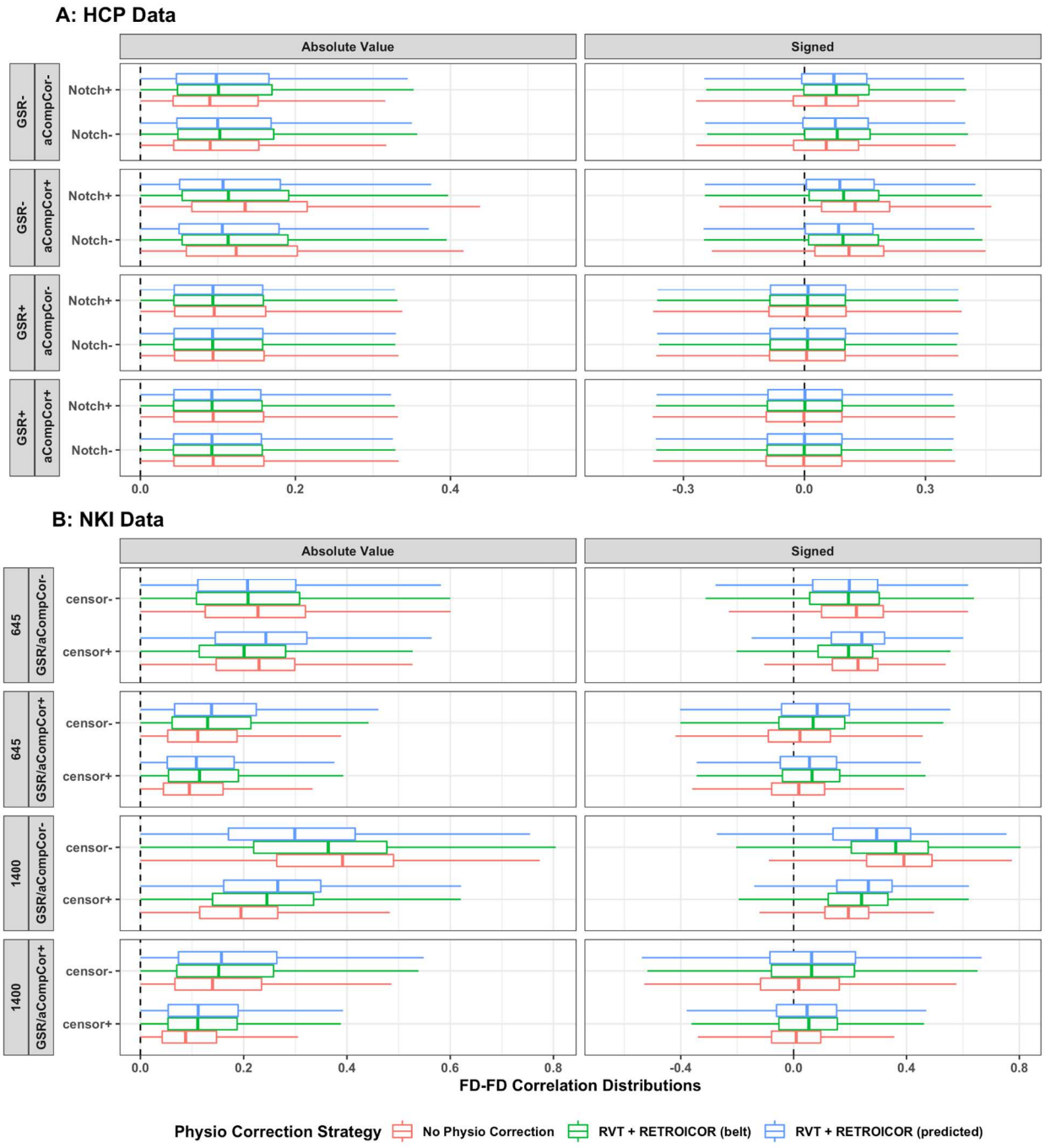


Figure 2.9. FD-FC correlations for all preprocessing specifications in the HCP (A) and NKI (B) datasets. Right panels show boxplots of distributions of signed FD-FC correlations (across all scans) between mean Jenkinson FD and functional connectivity for each edge (N=19900 edges for each boxplot). Left panels show boxplots of the absolute values of the same FD-FC distributions, representing the distribution of the magnitude of FD-FC correlations regardless of sign. For all panels, the y-axis labels indicate chosen forks for each pipeline (and within the NKI data, the sequence). For all plots, color indicates whether a model-based correction was used (red

= no physio correction, green = RVT + RETROICOR (belt), blue = RVT + RETROICOR (predicted). FD measurements based on raw BOLD data without any corrections were used for all FD-FC correlations.

Distance-Dependent FD-FC Correlations*:* We also examined how correction strategies impacted the distance-dependence of head motion artifacts by computing correlations between the length of edges (Euclidean distance between nodes) and corresponding head motion artifact (FD-FC correlation) for each edge. We then compared bootstrapped distributions of distance-dependent FD-FC correlations across pipelines. Across all pipelines and datasets, we observed negative correlations between edge length and FD-FC correlations, indicating that connectivity estimates between closer nodes were more positively associated with head motion across scans than were connectivity estimates between more distant nodes.

Within HCP data, neither notch filtering ($\Delta_{dist\_dependence}$ = -0.007, 95% CI [-0.031, -0.017], p = .5804), belt RVT + RETROICOR ($\Delta_{dist\_dependence}$ = 0.008, 95% CI [-0.021, 0.035], p = .5844), nor predicted RVT + RETROICOR ($\Delta_{dist\_dependence}$ = 0.012, 95% CI [-0.017, -0.041], p = .4266) impacted distance-dependent FD-FC correlations. However, GSR strengthened distance-dependent FD-FC correlations ($\Delta_{dist\_dependence}$ = -0.073, 95% CI [-0.096, -0.048], p = .0001), such that there was a stronger negative correlation between distance and FD-FC correlations with pipelines using GSR (Fig. 2.10). In particular, visualization of relationships between edge length and FD-FC correlations indicated that while GSR centered distributions of such correlations at 0, such correlations tended to be more negative for the longest edges (Fig. 2.10B & D). Conversely, aCompCor weakened distance-dependent FD-FC relationships ($\Delta_{dist\_dependence}$ = 0.037, 95% CI [0.014, 0.060], p = .0050).

RVT + RETROICOR using both belt ($\Delta_{dist\_dependence}$ = -0.059, 95% CI [-0.098, -0.020], p = .0001) and predicted ($\Delta_{dist\_dependence}$ = -0.078, 95% CI [-0.121, -0.036], p = .0001) traces

strengthened distance dependent FD-FC correlations within NKI-1400 data, though in the NKI-645 data neither RVT + RETROICOR with belt ($\Delta_{dist\_dependence}$ = -0.041, 95% CI [-0.094, 0.014], p = .1409) nor predicted ($\Delta_{dist\_dependence}$ = -0.039, 95% CI [-0.093, 0.014], p = .1389) traces had such consistent effects. The combination of both GSR and aCompCor did not impact distance-dependent FD-FC correlations in either the NKI-645 ($\Delta_{dist\_dependence}$ = -0.002, 95% CI [-0.046, 0.041], p = .9361) or NKI-1400 data ($\Delta_{dist\_dependence}$ = -0.020, 95% CI [-0.053, 0.137], p = .2088). Censoring weakened distance-dependent FD-FC correlations in both the NKI-645 ($\Delta_{dist\_dependence}$ = 0.071, 95% CI [0.027, 0.121], p = .0001) and NKI-1400 ($\Delta_{dist\_dependence}$ = 0.134, 95% CI [0.101, 0.166], p = .0001) data.

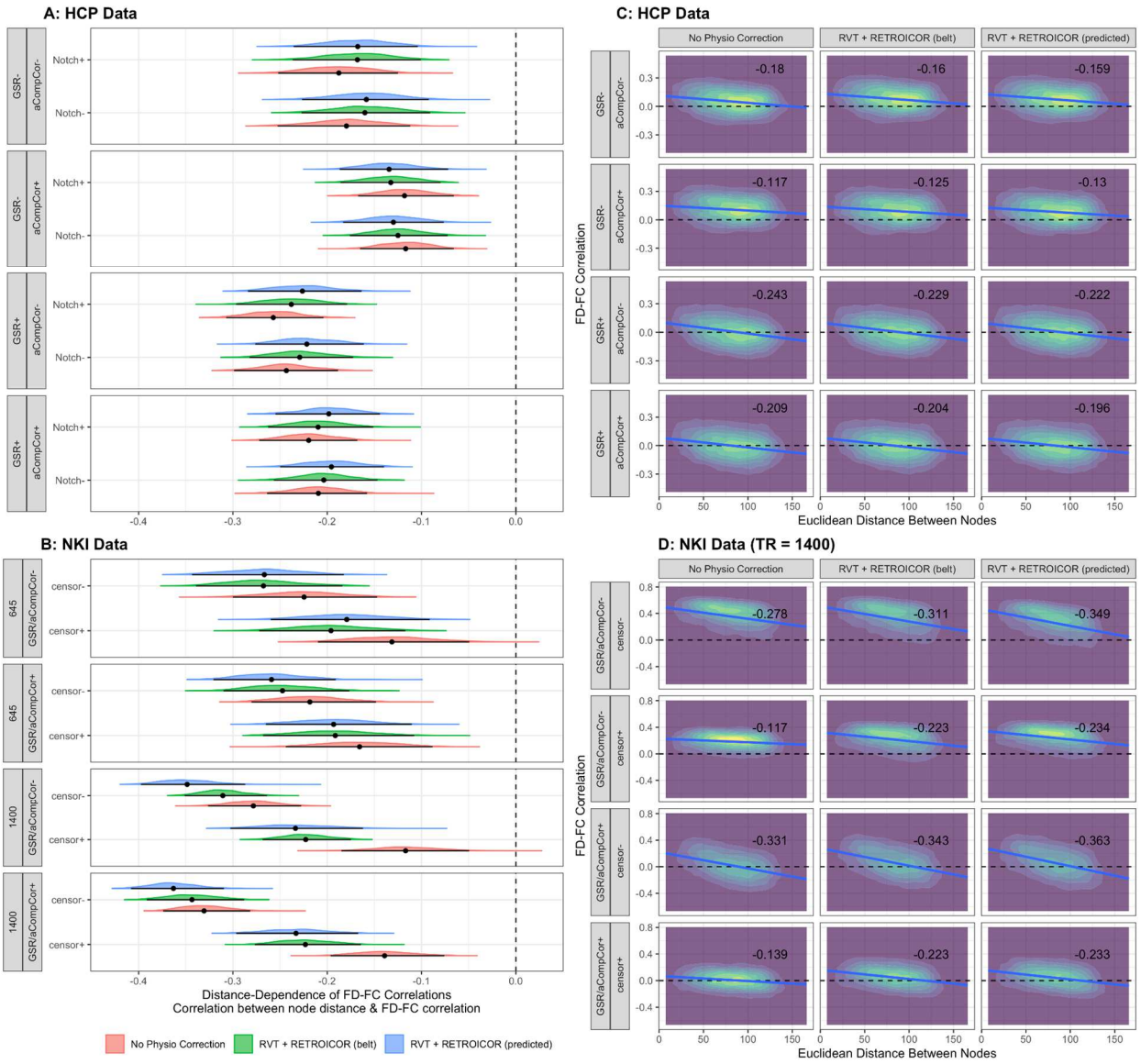


Figure 2.10. Distance-dependent FD-FC correlations. **A-B:** Bootstrapped distributions (using repeated random resampling of all scans, with replacement) of distance-dependence of FD-FC correlations for each pipeline for the HCP (**A**) and NKI (**B**) datasets. Distributions show rank-order correlations between edge length (Euclidean distance between nodes) and FD-FC correlations. Shaded densities indicate full bootstrapped distributions and points with error bars represent bootstrapped means and 95% confidence intervals. The dotted vertical line indicates zero linear distance-dependance. **C-D:** Bivariate distribution heatmaps showing Euclidean distance between nodes on the x-axis and FD-FC correlations for corresponding nodes on the y axis. Each subpanel indicates one pipeline as indicated by the row and column labels for the HCP (**C**) and NKI (**D**) datasets. Warmer, brighter color indicates higher density. Rank-order correlation lines of best fit are plotted over each heatmap, as well as the value of the mean

bootstrapped distance-dependent correlation coefficient. FD measurements based on raw BOLD data without any corrections were used for all distance-dependent FD-FC correlations.

Comparisons across QC Metrics*:* Because pipelines that perform comparatively better for one quality assurance metric may not necessarily perform better (and sometimes worse) on other metrics, we created an interactive web application for visualization of such potential tradeoffs (https://pbloom.shinyapps.io/qc_metric_comparison/). Here, users can compare the pipelines tested in the current investigation on metrics of data retention, test-retest reliability, and residual head motion artifacts, and visualize "tradeoffs" between optimizing on any pair of metrics in either the HCP or NKI data. For easier interpretability, only point estimates for each metric are displayed in the application (uncertainty estimates can be found in static manuscript figures).

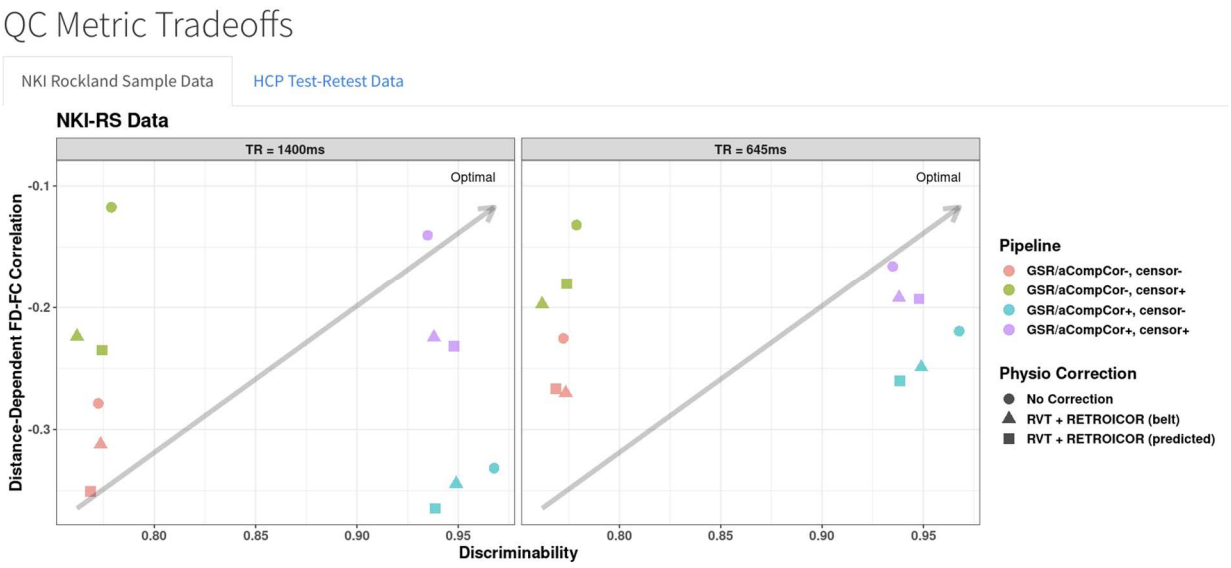


Figure 2.11. Interactive comparison of pipelines on combinations QC Metrics. Web-based application (https://pbloom.shinyapps.io/qc_metric_comparison/) allows for pipeline comparisons on any pair of QC metrics within the HCP & NKI data separately.

2.4 Discussion

Respiration-induced head motion is a major challenge for resting-state fMRI analyses, but has often been overlooked in investigations of preprocessing (Fair et al., 2020). Such artifacts are particularly visible in multiband datasets where data are acquired at higher frequencies, although often present in single-band data as well. Here, we attempt several different techniques for correcting respiration-induced head motion artifacts, and evaluate their impacts on data quantity, reliability, and presence of residual head motion artifacts. In particular, we compared techniques relying on peripheral collection of respiratory belt data or knowledge of participants' breathing rates (notch filtering head realignment parameters, RVT + RETROICOR with belt data) to techniques reliant only on BOLD (GSR, aCompCor, censoring, RVT + RETROICOR with predicted data). Building on prior work (Hocke & Frederick, 2021), we first demonstrate that corrections for breathing-induced artifacts can be performed using predicted respiratory traces generated from BOLD data alone, without requiring use of a peripheral respiratory belt. Next, we discuss the impacts of each of the techniques investigated on data quantity, reliability, and presence of residual head motion artifacts within both the HCP and NKI datasets.

A: HCP Data

Metric	GSR	aCompCor	Notch	RVT Pred	RVT Belt	GSR+, aCompCor+, RVT Pred **OR** Notch	GSR-, aCompCor+, RVT Pred **OR** Notch
Data Retention*	N/A	N/A	↑↑	↑↑	↑↑	↑↑	↑↑
Reliability	↑	↑				↑↑	↑

FD-FC	↓↓					↓↓	
FD-FC Distance-Dependence	↑↑	↓				↑	↓

B: NKI Data

Metric	GSR+aCompCor	Censor	RVT Pred	RVT Belt	GSR+, aCompCor+, Censor+, RVT Pred	GSR+, aCompCor+, Censor+
Data Retention*	N/A	↓↓	↑↑	↑	↑↑	↓↓
Reliability	↑↑				↑↑	↑↑
FD-FC	↓↓	↓	↑	↑	↓	↓↓
FD-FC Distance-Dependence		↓↓	↑	↑		↓

Table 2.3: Summary of impacts of correction strategies on QC metrics in the HCP (A) and NKI (B) datasets. Each cell indicates how a particular processing strategy (columns) impacted each of the QC metrics (rows). Upwards arrows indicate an increase in a given metric, and downward arrows indicate a decrease. Double arrows (↓↓, ↑↑) denote relatively larger effects compared to single arrows (↓, ↑). Green shaded columns on the right indicate pipelines that perform best on one or more metrics. 'RVT Pred' and 'RVT Belt' pipelines are shorthand for the combined RVT + RETROICOR step. Volume censoring was not conducted within the HCP data, and notch filtering was not conducted within the NKI data. *Note: if censoring is applied, pipeline choices impact data retention through reductions in framewise displacement. Data retention is not impacted if no censoring is done.

Prediction of respiratory traces from BOLD data: Using separate stack processing (Hocke & Frederick, 2021), we created "predicted" respiratory traces from high temporal resolution head motion estimates in the phase-encoding direction generated through separate stack processing. As previously reported, these predicted traces showed strong associations with corresponding respiratory belt traces in both the NKI and HCP datasets (Fig. 2.3). Correlations between belt and predicted respiratory traces were often strongest with a slight (1-2.5s) lag, indicating potential phase delay between the belt and predicted traces (Fig. 2.3, Appendix B Fig. 6), although correlations were still present without phase delay. Additionally, peak frequencies were highly similar between belt and predicted traces, indicating that such predicted traces could capture frequency information measured by the respiratory belt.

Although the current results within the NKI retest session data were quite similar to those previously reported in a distinct subsample of NKI scans (Hocke & Frederick, 2021), our findings indicated somewhat weaker relationships between predicted and belt respiratory traces. A potential reason for this is that the current study did not employ strict quality control-based inclusion criteria for either the belt traces or BOLD data, and NKI data in the current investigation included both BOLD runs with high levels of head motion and low-quality respiratory traces (for example, high rates of signal clipping or evidence of the belt slipping off). Among the NKI data, correspondence was indeed weaker between belt and predicted traces when either belt or BOLD data quality was lower (see Figs.s 5-6). In particular, that scans with lower mean framewise displacement showed stronger similarity between belt and predicted traces indicates that high levels of head motion may impede estimation of respiratory information from BOLD data.

Data Inclusion*:* In both HCP and NKI-RS datasets, model-based respiratory correction (RVT + RETROICOR) using predicted respiratory traces mitigated respiratory artifacts within head motion estimates, and reduced the amount of data excluded when censoring based on head motion comparably to a notch filtering strategy (Fair et al., 2020). One drawback to the notch filter observed in the HCP data in particular was a somewhat higher proportion of "lost" volumes that were included without the correction, but above threshold for exclusion afterwards. As has been previously noted within the context of BOLD data itself (Power et al., 2013), such band-pass filtering can "spread" impacts of high-motion frames to neighboring ones. It is possible that notch filtering the head realignment parameters is producing the same effect and spreading motion estimates from high-motion frames to ones immediately before or after.

Although data quantity is an essential component of statistical power and generalizability (Chen et al., 2022; Cho et al., 2020; Gordon et al., 2017; Marek et al., 2020; Nee, 2019), we emphasize that increased data quantity alone cannot be a benchmark for optimization. In particular, relaxing motion-based inclusion criteria without simultaneous verification of data quality risks increased contamination by head motion artifacts. In particular, recent work has suggested that conventional FD=0.2mm thresholds may need to be lowered after applying correction methods that broadly decrease the power of the FD trace (Gratton et al., 2020; Kaplan et al., 2022). Therefore, increases in data retention after notch filtering the head realignment parameters or under RVT + RETROICOR approaches may not be beneficial unless equivalent or better data quality can be established.

Reliability: Overall, the current reliability findings suggest that while corrections for breathing-induced artifacts may impact head motion, their effects on preprocessed whole-brain functional connectivity estimates are dwarfed by those of GSR and aCompCor (Xifra-Porxas et

al., 2021). In particular, notch filtering did not impact inter-pipeline agreement, I2C2, or discriminability. RVT + RETROICOR, whether using belt or predicted respiratory traces, only had minor impacts on inter-pipeline agreement, and did not impact I2C2 or discriminability. Although RVT + RETROICOR tended to decrease between-participant and within-participant variance consistent with previous findings (Birn et al., 2014a), effects were small and did not impact I2C2 or discriminability.

In contrast, GSR and aCompCor had major impacts on inter-pipeline agreement, as well as reliability between sessions and sequences. GSR in particular impacted inter-pipeline agreement the most (Li et al., 2021). GSR and aCompCor both contributed to large reductions in between-participant and within-participant variance, which lead to increases in discriminability. While neither step consistently impacted I2C2, lack of convergence between the I2C2 and discriminability metrics may be due to both violation of the gaussian assumptions of the I2C2 metric and differences in their calculation. Whereas I2C2 is a ratio of between-participants to within-participants variances, discriminability is calculated as the average proportion of within-participant distances that are smaller than between-participant distances (Li et al., 2021; Milham et al., 2021).

While reliability is critical for fMRI measurement (Elliott et al., 2021), reliability alone is not sufficient for pipeline optimization. Highly reliable signal is not necessarily valid signal, and trait-like motion or respiratory artifacts in functional connectivity can boost reliability while contaminating neuronally-based signals (Birn et al., 2014a; Finn & Rosenberg, 2021; Siegel et al., 2014). To ensure that this is not the case, pipelines can be optimized both for reliability and reduction of such artifacts.

Residual head motion artifacts: To compare residual head motion artifacts in functional connectivity across pipelines, we computed FD-FC correlations and their distance dependence. As previously reported (Ciric et al., 2017b; Power et al., 2014), GSR reduced FD-FC correlations overall while increasing the distance-dependence of FD-FC correlations, and censoring reduced both overall and distance-dependent FD-FC correlations. While GSR worsened distance-dependent artifacts, simultaneous use of aCompCor partially mitigated this issue in the HCP data. Although we did not test GSR and aCompCor factorially in the NKI data, pipelines with both GSR and aCompCor showed roughly equivalent distance-dependence to pipelines with neither step. Decisions for whether to include GSR in preprocessing may then depend on the relative priority of minimizing the central tendency of FD-FC relationships versus their distance-dependence, although aCompCor and censoring help improve both metrics (see Table 2.3).

While findings indicated that while RVT + RETROICOR may benefit data retention, including such data may come at the expense of exacerbated residual head motion artifacts in functional connectivity estimates. However, such effects were specific to the NKI data, such that neither belt nor predicted RVT + RETROICOR impacted FD-FC correlations or their distance dependence in the HCP data. It is possible that such effects in the NKI are due to the fact that data quality was lower (higher motion, clipping of respiratory belt traces) in some of the NKI scans. In such scans, conducting RVT + RETROICOR before rigid body motion correction of the fMRI data may have interfered with this step, as well as the estimated head realignment parameters, causing downstream worsening of FD-FC relationships. Future work could examine whether integrating RVT + RETROICOR with rigid body correction effectively reduces these FD-FC associations (Jones et al., 2008).

It is possible that RVT + RETROICOR may have led to the inclusion of more motion-contaminated volumes, which in turn worsened FD-FC relationships. The fact that the RVT + RETROICOR did not impact FD-FC relationships in the HCP data, where we did not apply motion-based censoring, supports this possibility. However, in the NKI data, increased FD-FC artifacts were still present in pipelines with either belt or predicted RVT + RETROICOR even when censoring was not done. Therefore, the mechanism by which such model-based respiratory corrections may worsen residual head motion artifacts is yet unclear.

Model-based versus frequency-based respiratory correction strategies: Current findings indicate that researchers have numerous viable options for mitigation of breathing-induced artifacts even without available respiratory belt data. However, because thus far the current study has not examined impacts of notch filtering on functional connectivity in the NKI dataset or censoring in the HCP dataset, conclusions cannot be made on the relative strengths and weaknesses of notch filtering versus predicted RVT + RETROICOR. While such comparisons will be crucial for more definitive recommendations, we summarize relevant findings and considerations so far below.

First, model-based strategies have several practical and theoretical advantages over notch filtering. Most notably, notch filtering is only possible in datasets with sufficient temporal resolution to estimate head motion in the 0.2-0.4Hz frequency range (data would need to be collected at TR=1.25 to estimate motion up to 0.4Hz). However, low-pass filtering may be used at slower TRs (Gratton et al., 2020), though this strategy has not been investigated in depth in the current investigation. In addition, RVT + RETROICOR is tailored to mitigate participant-specific and scan-specific breathing artifacts (Birn et al., 2014b), whereas static notch filtering is inflexible to differences in respiration patterns between participants (if a common filter is applied

to a whole dataset) or within the course of a scan. Particularly in datasets where breathing rates may vary widely (i.e. datasets with wide age ranges), designing filters that capture respiratory frequencies for most participants could be difficult (Fair et al., 2020). Yes, within the HCP data, such theoretical advantages of model-based correction approaches did not result in improved reliability (Fig. 2.7-8) or reduced residual head motion artifacts (Fig. 2.9-10).

At the same time, several key factors may make notch filtering head realignment parameters preferable over model-based strategies. Notch filtering is a more parsimonious and less computationally demanding strategy compared to the slice-based regressions used under RVT + RETROICOR, and is a more accessible tool for researchers given its implementation in existing software (Craddock et al., 2013; Fair et al., 2020). Further, previous work has indicated that this approach can increase the reliability of functional connectivity estimates (Fair et al., 2020; Kaplan et al., 2022), although current findings did not confirm this in the HCP data. Finally, we found that RVT + RETROICOR using belt or predicted traces increased residual FD-FC relationships in the NKI data (see Fig. 2.9-10). If further analyses in the NKI data indicates that such artifacts are not increased by a notch filter, this would be another piece of evidence for its relative advantage.

At present, however, findings do not support strong guidelines for choosing between model-based and filter-based approaches to respiratory artifact correction. For now, we suggest that either strategy may be applied in conjunction with other tools, depending on one's priority in quality control metrics.

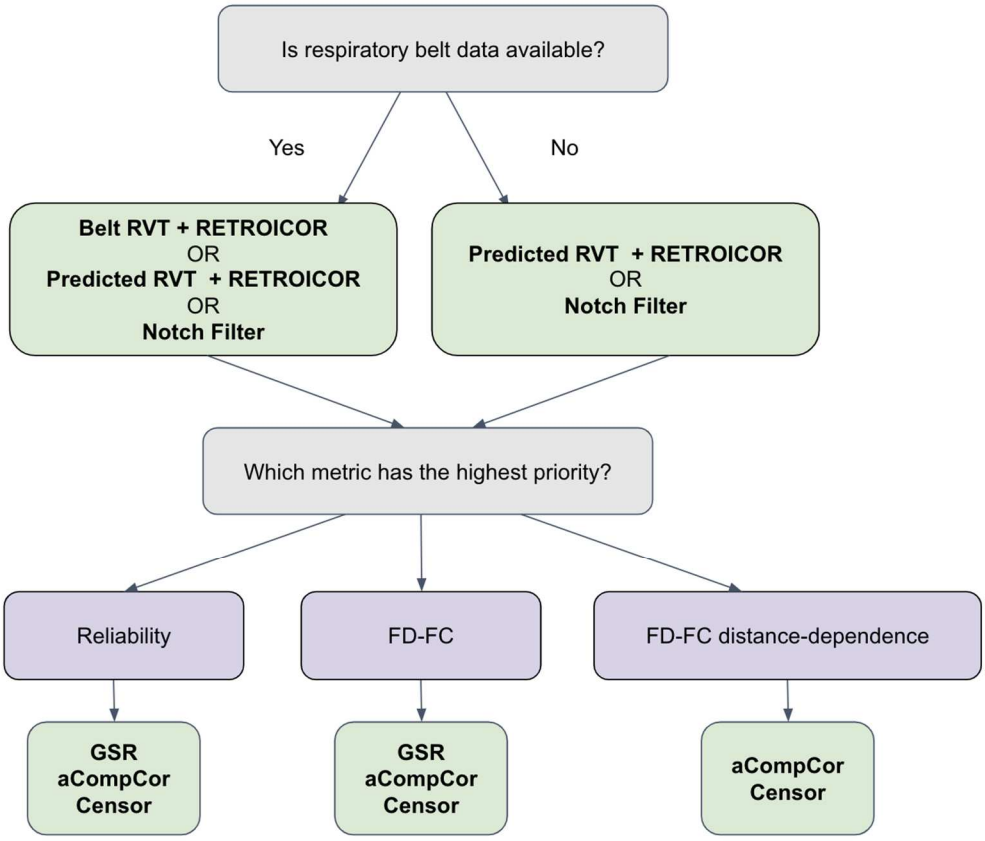


Figure 2.12: Decision tree for guiding preprocessing choices. Optimal pipelines may depend on one's priority in quality control metrics.

Summary of Observations: Overall, our findings indicated that no one pipeline optimized all metrics tested. Although aCompCor improved both reliability and reduced residual head motion artifacts with few drawbacks, all other individual methods represented "tradeoffs" between two or more metrics. In particular, while predicted RVT + RETROICOR demonstrated viability for reducing breathing-induced artifacts in head motion estimates, this strategy came at the expense of exacerbated head motion artifacts in functional connectivity in a developmental (ages 6-20) dataset (NKI). In efforts to guide researchers in selection of pipelines to meet their priorities in quality control metrics, we summarize the impacts of each of the tested strategies

(Table 2.4) and construct a decision tree for potential ways to approach such decisions (Fig. 2.12).

Topic	Main Observations
Notch filtering head motion parameters *Functional connectivity HCP only	1. Reduced respiratory artifacts in head motion estimates, which improved data retention when censoring based on head motion 2. Did not impact reliability inter-pipeline agreement of functional connectivity 3. Did not impact head motion artifacts in functional connectivity or their distance-dependence
RVT + RETROICOR	1. Reduced respiratory artifacts in head motion estimates, especially when using predicted traces, which can improve data retention comparably to notch filtering 2. Did not impact reliability of functional connectivity when either belt or predicted traces were used, and only had minor impacts on inter-pipeline agreement 3. Increased head motion artifacts in functional connectivity among pipelines including GSR and aCompCor, though such effects were small in magnitude compared to those of GSR (NKI only) 4. Increased distance-dependent head motion artifacts in functional connectivity, though such effects were small in magnitude compared to those of censoring (NKI only)
GSR	1. Increased discriminability, but not I2C2. 2. Impacted inter-pipeline agreement more than any other single step 3. Reduced both within-participant and between-participant variances in functional connectivity 4. Reduced head motion artifacts in functional connectivity, but increased their distance-dependence
aCompCor	1. Increased discriminability, but not I2C2 2. Impacted inter-pipeline agreement less than GSR, but more than any other steps 3. Reduced both within-participant and between-participant variance in functional connectivity 4. Reduced distance-dependent head motion artifacts
Censoring *NKI only	1. Reduced data retention (by definition), but did not impact reliability 2. Reduced both residual head motion artifacts and their distance-dependence
General QC Points	1. Given the same threshold (such as FD > 0.2mm), using the Power FD metric will always result in equal or more data excluded compared to the Jenkinson FD metric 2. aCompCor benefited several QC metrics and showed few negative impactsMultiple QC metrics should be considered in tandem with the characteristics of the scanned cohort and research question when making fMRI preprocessing decisions 3. Investigations of reliability should separately examine within-participant and between-participant variance

	4. Resting-state fMRI studies will be strengthened by multiversing key preprocessing choices, most notably GSR and corrections for breathing-induced artifacts in head motion (Li et al., 2021)

Table 2.4: Summary of observations for each preprocessing strategy, as well as general quality control considerations

Limitations: We note several key limitations to the present work. Most notably, distinguishing breathing-induced pseudomotion from true head motion (breathing-induced or otherwise) remains a challenge. Although use of multi-echo sequences may help with discrimination of artifact signals in BOLD data (Kundu et al., 2015; Power et al., 2018), future fMRI denoising work will also benefit from "ground truth" peripheral measurements of head position not susceptible to breathing-induced distortions. Unfortunately, that respiratory correction techniques, particularly RVT + RETROICOR, may be most effective in high-quality low-motion data limits their effectiveness in contexts where exclusion criteria may be most important, including developmental or multi-site investigations.

An additional limitation is that the current study did not test many possible preprocessing strategies for mitigating both motion and breathing-induced artifacts. Impacts of notch filtering on functional connectivity were not tested in the NKI data, and impacts of censoring not tested in the HCP data, which makes weighing the relative strengths and weaknesses of predicted RVT + RETROICOR and notch filtering across datasets more difficult. While exhaustive examination of all possible strategies is beyond the scope of any one investigation, future work could address in particular whether scan-specific notch filtering strategies (potentially using predicted respiratory traces to select filter envelopes) or independent components analysis (ICA-FIX, ICA-AROMA; (Pruim et al., 2015) effectively mitigate such artifacts. Further, the current work did not directly test whether high-resolution estimation of head motion parameters can be used to predict respiratory signal in single-band fMRI datasets, where breathing-related head motion artifacts are

also present (Gratton et al., 2020). In theory, this may be feasible if head realignment parameters are estimated separately for individual slices (rather than "stacks" of slices), though further investigation will be needed to explore this possibility.

Finally, we note several limitations to some of the QC metrics used in the current study. Since both I2C2 and discriminability are multivariate metrics of reliability, findings based on the multivariate reliability of functional connectivity matrices as a whole may not necessarily apply to individual edges or regions (Xu et al., 2022). In addition, recent work has demonstrated that FD-FC correlations may be an imperfect metric (Raval et al., 2021), as addition of high-variance noise in functional connectivity can suppress true associations with head motion. Last, the current investigation examined impacts of preprocessing strategies on static functional connectivity during resting-state fMRI. Thus, findings may not generalize to task-based approaches, scans collected during movie-watching (Vanderwal et al., 2019), or time-varying connectivity methods (Bassett et al., 2011).

Conclusion

Broadly, the current findings indicate necessary tradeoffs between data inclusion, data reliability, and residual head motion artifacts in resting-state fMRI preprocessing, as no one pipeline was able to optimize all such metrics. Given such tradeoffs, future studies will also likely benefit from tailoring pipelines to match priorities in quality control metrics, and "multiverses" of preprocessing pipelines to determine the robustness of results to potentially influential analytical decisions (Botvinik-Nezer et al., 2020; Cosme & Lopez, 2020; Dafflon et al., 2020).

Acknowledgements

Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Chapter 3: Strong and wrong? An exploration of strategies beyond model-free and model-based learning from childhood to young adulthood

Paul Alexander Bloom, Andrea Fields, Tricia Choy, Nicolas Camacho, Lisa Gibson, Anna Vannucci, Syntia Hadis, Chelsea Harmon, Sage Hess, Gabriela Rodrigues, Michael Lincoln, Roxanna Flores Toussaint, Rebecca Umbach, Charlotte Heleniak, Daphna Shohamy, & Nim Tottenham

Abstract

Recent work has suggested that the ability to plan prospectively using cognitive maps of the environment undergoes maturation from middle childhood to young adulthood. While adolescents and adults use such "model-based" learning, children have been argued to use "model-free" strategies without planning ahead or using environmental structures to make decisions. Such findings have been replicated across several studies using two-stage sequential decision-making paradigms. Yet previous analyses of developing two-stage task behaviors have rarely studied decision-making at the second stage of the task. Here, we studied two-stage task strategies in 62 youth ages 7-13 years old using a modified two-stage task intended to enhance model-based decision making. Contrary to expectations, there was no evidence for the signatures of model-based or model-free decision-making during the first stage of the task. Yet, child behaviors indicated sensitivity to task structure. To better understand such behaviors, we compared data collected under this modified paradigm to three previously collected developmental datasets (301 total participants, ages 8-25 years). Despite methodological differences (e.g., task instructions, gamification, trial duration, reward incentives, visit duration) across datasets we also observed behavioral patterns at the second stage of the task explained neither by model-free nor model-based algorithms. Participants across the entire age range studied repeated *spatial-motor sequences* (left/right button presses) following rewards, even when such sequences would not impact the likelihood of receiving a subsequent reward. Further, participants repeated such sequences more quickly following rewarded trials, suggesting that such behaviors may be planned before the start of the subsequent trial. These findings suggest that, in tandem with model-free or model-based strategies, youth assign value to spatial and motor cues in environments where the outcomes of their decisions are uncertain.

3.1 Introduction

Starting even before birth (Spence & DeCasper, 1987; Varendi et al., 1996; Voegtline et al., 2013), learning from one's environment to guide future behavior is a vital feature of human development (Knudsen, 2004; Rovee & Rovee, 1969). The ability to use previous positive and negative experiences to choose actions is necessary for adaptive functioning across development, including both fundamental regulatory needs and higher order social and cognitive processes (Frankenhuis et al., 2019; Jones et al., 2011; Nussenbaum & Hartley, 2019). In particular, *reinforcement learning,* or the association of prior actions with outcomes and subsequent use of associations to choose future actions, is fundamental to learning how to successfully navigate both threatening and rewarding environments (Delgado et al., 2008; LeDoux & Daw, 2018). Reinforcement learning algorithms (Rescorla & Wagner, 1972; Sutton & Barto, 1998) drawn from computer science have shown efficacy in explaining both human and animal behaviors during reward (O'Doherty et al., 2003; Roesch et al., 2012) and threat learning (Mkrtchian et al., 2017; Phelps et al., 2004), as well as corresponding midbrain dopaminergic signals (Schultz et al., 1997; Sharp et al., 2016). Thus, reinforcement learning paradigms have been argued to have both explanatory potential for a variety of behaviors and biological plausibility.

Recently, much work has focused on distinguishing behaviors generated by two distinct reinforcement algorithms, dubbed "model-free" and "model-based" learning (Daw et al., 2011; Gläscher et al., 2010). Model-free learners associate actions with values, but do not use "cognitive maps" of the environment to link actions with subsequent states (a state can be any new situation with a new set of possible actions). Thus, model-free learners only retrospectively update the utility of actions, and do not "plan ahead". Model-based learners, on the other hand, make prospective decisions based on "maps" of the environment, which include both values of

actions and the likelihood that actions will lead to future states (Doll et al., 2012, 2015; Duncan et al., 2018). For example, while a model-free learner might learn from previous trips which of several routes of travel tends to be fastest, a model-based learner might also learn which routes tend to be fastest conditional on the time of day, then plan their trip based on what time of day they will be leaving.

While model-based learning is more capable of nuance and likely advantageous in many environments, such a strategy is thought to come at cognitive burden. Model-based learners must store and update more information (i.e. both maps and action values, whereas model-free learners only keep track of action values), especially in complex environments (Otto, Raio, et al., 2013). Thus, empirical research and simulation studies have indicated that it may be adaptive to use a combination of both strategies, or switch between strategies as a function of one's environment (Simon & Daw, 2011). Indeed, adult humans may use combinations of both model-based and model-free learning (Daw et al., 2011), or even change strategies over time. Nevertheless, whether model-based versus model-free approaches are more computationally demanding, from an algorithmic perspective, depends on the structure of the environment. Indeed, environments exist where model-free strategies are no less costly than model-based ones (Kool et al., 2016; Simon & Daw, 2011). Further, the "cost" of a strategy for an algorithmic agent (number of required computational steps, parameters required, values cached) may not map directly onto required memory and executive functions for human participants.

In particular, research with human participants has approached studying these learning strategies with versions of a 'two-stage' Markov decision task, which assess use of both types of learning (Daw et al., 2011; Kool et al., 2016). In the most widely-used paradigm (Daw et al., 2011), participants navigate through two stages of sequential choices over many trials. In the first

stage, participants make a binary choice of two stimuli, which each lead probabilistically (typically, each of the two stage 1 choices lead to one of the stages with a 70% likelihood, and a 30% likelihood to the other) to one of two second-stage "states". In each of the two states, participants then make a binary choice between two more stimuli, each of which provide a reward with a different likelihood. The reward probabilities of each of the four stage 2 choices drift independently over time according to a Gaussian random walk procedure. Typically, researchers examine patterns of choices at stage 1 over several hundred trials to ask whether choices at this stage reflect use of the "transition probabilities" from the first to second stages in decision-making. Participants using model-based strategies incorporate the transition probabilities into decision-making, making stage 1 choices with the plan to reach a specific state at stage 2, while participants using model-free strategies use a simpler strategy of repeating stage 1 choices that previously lead to rewards (regardless of stage 2 state, see Fig. 3.2).

Model-based and model-free learning have also been proposed as a useful framework for understanding cognitive development from childhood to young adulthood (Raab & Hartley, 2018). In particular, recent work has held that model-based learning is a hallmark of cognitive development, as use of model-based learning strategies tend to increase from childhood through young adulthood (Decker et al., 2016; Nussenbaum et al., 2020; Potter et al., 2017). In particular, youth ages 13 years and under have been argued to rely mostly on "habit', or model-free strategies. Further, model-based learning is associated with abstract reasoning capabilities in developmental samples (Nussenbaum et al., 2020; Potter et al., 2017), motivating arguments that the developmental appearance of model-based learning may be driven by the slower maturation of prefrontal cortical (PFC) function compared to the basal ganglia (Casey et al., 2016). Supporting this view, disruption of the dorsolateral PFC through transcranial magnetic

stimulation decreased model-based, but not model-free learning in adults (Smittenaar et al., 2013).

On the other hand, there is ample evidence that infants can learn hierarchical rule structures (Werchan et al., 2015, 2016) and apply them to new stimuli. Such behaviors are closely related to model-based control, as they require agents to learn context-dependent rules and choose actions accordingly (Frank & Badre, 2012). Further, work with a modified two-stage task paradigm has found that children as young as 5 years of age are capable of model-based control (Smid et al., 2020). Across several studies, children ages 8-13 also demonstrate *knowledge* (if not use of) of transition structure from stage 1 to stage 2 within the two-stage paradigm as evidenced by increased reaction times following "rare" compared to "common" transitions (Nussenbaum et al., 2020). Together, such findings raise the possibility that younger children may elect to use strategies other than model-based control to approach the two-stage task, rather than lacking the neural circuitry to support such computations.

Simultaneously, recent work has demonstrated that adult participants' behaviors during two-stage task paradigms are heavily influenced by task instructions (Feher da Silva & Hare, 2020). In particular, if task instructions are not sufficiently clear, adult participants make assumptions about the structure that may bias estimates of model-free versus model-based learning. Recent work has argued that adults primarily use model-based strategies, and apparent combinations of model-free and model-based strategies are due to underspecified task instructions or improper analysis strategies (Miller et al., 2016). Modifying the task with more specific instructions in one study led to little evidence of model-free learning (Feher da Silva & Hare, 2020). Thus, it may be possible that age-related differences in interpretations of

instructions may give rise to findings of increasing model-based control between childhood and young adulthood.

Further, the continuum from model-free to model-based learning is only one dimension with the potential to explain behavior in a potentially multidimensional space of task strategies (Collins & Cockburn, 2020; Momennejad et al., 2017). Particularly because "credit assignment" (e.g. determining what contributed to a reward; Minsky, 1961) following stochastic rewards is difficult, participants may also assign value to reward-independent spatial or motor information during such two-stage paradigms (i.e. repeating patterns of left/right button presses after rewards, regardless of transitions and states), rather than pursuing putatively model-based policies (Shahar et al., 2019). However, little work has investigated strategies beyond purely model-free and model-based ones during the two-stage paradigm within developmental samples. Despite the fact that there is much active research using this paradigm, deeper knowledge of the degree to which additional learning mechanisms may help to accurately characterize behavior would be highly beneficial for understanding the development of reward processing.

In the present work, we take several approaches to more holistic characterization of youths' learning strategies during two-stage paradigms. Initially, we designed a modified space-themed version of the two-stage task intended to encourage model-based learning through a constant spatial representation of the transition structure. Healthy developing children (N=62, ages 7-13 years) completed this task in the laboratory. Contrary to our expectations, participants showed neither typical model-free nor model-based signatures at the first stage of the task. In efforts to understand such strategies, we conducted exploratory analyses of stage 2 behaviors under our modified paradigm, as well as in three previously collected datasets. Overall, while results indicated behaviors unique to the current paradigm at stage 1, we found evidence for

stage 2 behaviors in all datasets not explained by either model-free or model-based learning algorithms. In particular, participants across the entire age range studied repeated *spatial-motor sequences* (left/right button presses) following rewards, even when such sequences did not impact the likelihood of receiving a subsequent reward (Shahar et al., 2019). Overall, the current findings indicate that model-free and model-based signatures may be both sensitive to both the task (instructions, practice trials, spatial layout of stimuli, trial duration, 'gamification') and study design (monetary incentivization of rewards, context of task within study session) factors. Yet, the presence across paradigms of credit assignment to spatial-motor cues indicated a broader tendency to use such cues even when independent of reward outcomes.

3.2 Methods

Participants: The present study included 62 (30M / 32F) youth recruited from the United States ranging from 7-13 years of age (M=9.6, SD = 1.74). The sample studied here was part of the Parents and Children Coming Together (PACCT) study (Fields et al., 2021; Nikolaidis et al., 2022); a larger longitudinal study on early caregiving disruptions (N = 103, institutionalization, domestic or international foster care, extended separation from parents), though the present work focused only on "comparison" youth without such experiences. The median family income-to-needs ratio for participating families was 2.35 (SD=3.4, range [0.44, 14.8]). Information on participant race and ethnicity is available in supplemental tables 1-2. Participants studied in the current work were recruited via flyers, street fairs, word of mouth, and re-contact from lists of interested participants from previous lab studies. Data collected for the current work were from the second visit of the longitudinal study, which occurred on average 18 months after the first visit. Parents provided consent and youth provided written assent.

Study visits most often lasted 4-5 hours, and included approximately 75 minutes of magnetic resonance imaging (MRI). Most youth completed MRI scanning prior to the two-stage sequential learning task studied here. In addition, youth completed a larger set of assessments aimed at characterizing cognitive control behavior and negative valence systems, while parents completed surveys and semi-structured clinical interviews (KSADS). Analyses in the current study only include the two-stage learning task data, and not the MRI or other assessments. The university Institutional Review Board approved the study protocol, and families were compensated $175 per child (for some families, multiple children participated) in cash. Travel expense coverage (Uber for families within New York City, Amtrak tickets and hotel vouchers for families outside New York City) was also offered to participating families.

Data collection of the two-stage task began in March 2019 and was ended in March 2020 at the onset of the COVID-19 pandemic. Although moving the task paradigm to an online format was possible (Nussenbaum et al., 2020), we chose to stop data collection given that a majority of healthy comparison participants (62/106, ~58%) had already completed the task, and to avoid discrepancies in procedures between participants.

Previously Collected Datasets: In addition to data collected specifically for the present study, we conducted secondary data analyses of 3 previously acquired datasets using the spaceship version of the two-stage task from which the current paradigm was based (Decker et al., 2016; Nussenbaum et al., 2020; Potter et al., 2017). All data were accessed via the Open Science Framework (https://osf.io/we89v). All three datasets used similar versions of the spaceship task, and included participants ranging in age from 8-25 years.

Source	Name used here	Age Ranges (years)	Collection Method

Decker et al., 2016	Decker dataset	8-12 (N=30), 13-17 (N=28), 18-25 (N=22)	In-lab
Potter et al., 2017	Potter dataset	8-12 (N=26), 13-17 (N=23), 18-25 (N=25)	In-lab, during fMRI
Nussenbaum et al., 2020	Nussenbaum dataset	8-12 (N=50), 13-17 (N=50), 18-25 (N-51)	Online via Pavlovia

Table 3.1: Information on three previously collected developmental datasets using a spaceship version of the two-stage task

Two-Stage Task: Participants completed a custom version of the two-stage task built in pygame (https://github.com/pab2163/spaceTreasureRLTask) and inspired by the one used by Decker et al. (2016). The task structure was identical, such that participants were presented with 200 trials each where they first selected a green or yellow "portal" (Fig. 3.1A) to move their avatar using the left/right arrow keys. Then (Fig. 3.1B), on 70% of trials, a ladder carried the avatar vertically to the planet above (common transition), where in the other 30% of trials a ladder carried the avatar diagonally to the other planet (rare transition). Once at a given planet, the participants then used the left/right arrow keys again to select an alien to approach (Fig. 3.1C). On each trial, aliens gave coin rewards at likelihoods that drifted independently using a Gaussian random walk (with standard deviation 0.025) over the course of the task (Fig. 3.1D). Reward likelihood changed over the course of the task in order to incentive learning over all 200 trials (Daw et al., 2011). The positions of the different portals and aliens were counterbalanced across participants, but fixed within participants such that neither portals nor aliens shifted in location for a given run. The task was designed such that exactly 7 of the first 10 trials, in a random order, would be common transitions (so as not to bias participants' cognitive maps early in the task). Then 133 of the 190 remaining trials were set to be common transitions, also in a random order.

While the core structure of the task design was identical to that used in previous studies (Daw et al., 2011; Decker et al., 2016; Nussenbaum et al., 2020; Potter et al., 2017), we also made several design modifications intended to enhance participant engagement and encourage model-based learning strategies. Most notably, the entire structure of both stages 1 and 2 remained on the screen at all times, such that participants could see both planets and all four aliens even when at stage 1, as well as seeing the chosen portal and ladder while at stage 2. We hoped that by making the spatial transition structure always visible to participants, this would increase *use* of this structure for model-based learning.

In addition, because participants were completing this task during the middle of a 4-hour study session, we made several modifications to enhance engagement with the task. First, we increased the speed of transitions to 500ms, and decreased the intertrial interval to 500ms such that each trial would take ~3s and the entire task around 10 minutes. We also allowed participants to choose their own avatars from a set of possibilities (Black Panther, Chloe the cat [from *Secret Life of Pets*], Spiderman, or Pikachu), and included sound effects and background music throughout the task in efforts to resemble a video game. To further emphasize reward feedback, the game emitted a "whee!" sound when the alien choice resulted in a coin reward, and a beeping sound when the choice did not result in a reward. Finally, we shortened the instructions and practice period such that participants viewed an overview of the task structure, then completed only 2 practice trials with a different stimulus set. This was substantially shorter compared to the 50 practice trials completed by participants in other studies (Daw et al., 2011; Nussenbaum et al., 2020). A video of the tutorial and 25 trials of the task can be viewed at https://osf.io/9m8xh/. Due to the structure of the larger study session, participants completed the task in a mock MRI room or participant testing room on the same floor as the MR suite on a

Windows laptop with the experimenter sitting next to them. Participants were not in the MRI scanner while completing the task.

After completion of the task, we also asked a subset of participants to report "what strategies did you use to play the game?" in their own words. Experimenters transcribed these responses word-for-word as accurately as possible.

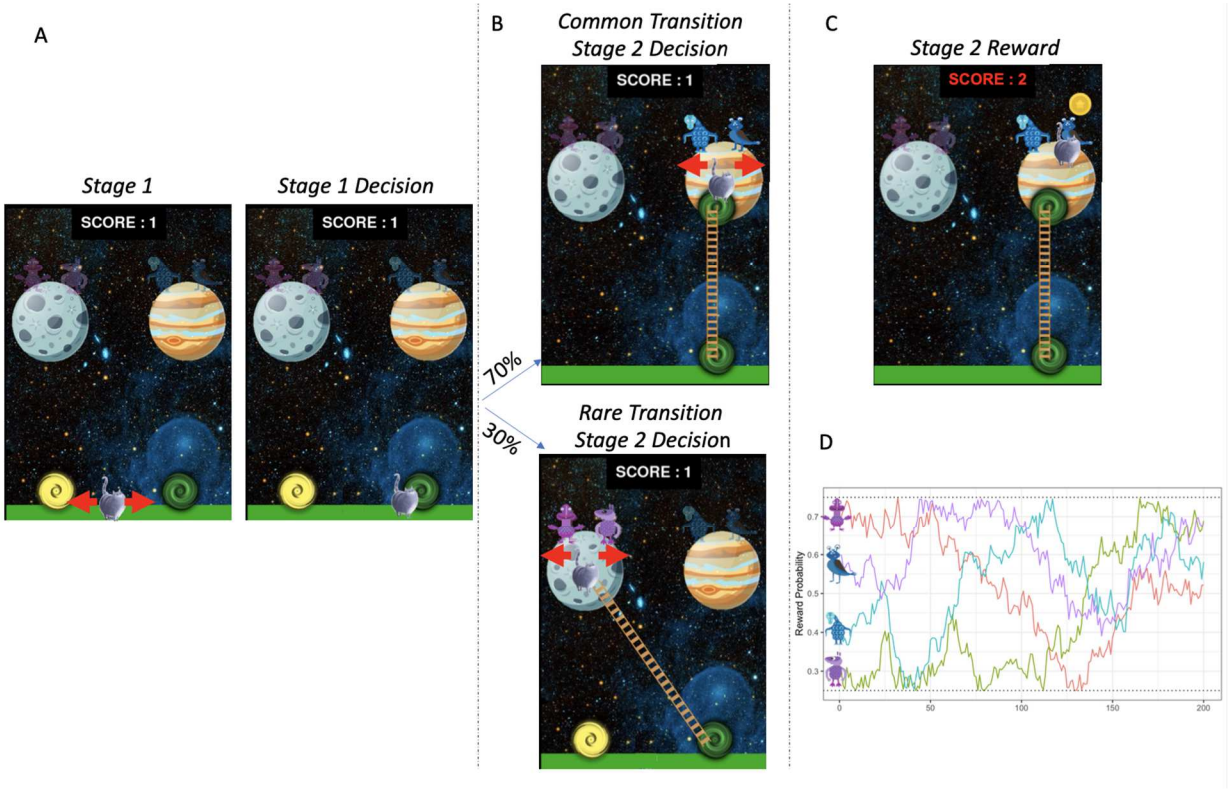


Figure 3.1: Participants completed a modified version of the space-themed two-stage task developed by Decker et al. (2016). For each of 200 trials, participants first selected a portal (A) to navigate their avatar using the left/right arrow keys. Then (B), on 70% of trials, a ladder carried the avatar vertically to the planet above (common transition), where in the other 30% of trials a ladder carried the avatar diagonally to the other planet (rare transition). Once at a given planet, the participants then used the left/right arrow keys again to select an alien to approach (C). On each trial, aliens gave coin rewards at likelihoods that drifted independently over the course of the task (D). Positions of all portals, planets, and aliens were counterbalanced between participants but were constant across trials for each participant. Portals, planets, and aliens were

always visible on the screen to encourage use of the transition structure for decision-making, although aliens were slightly more transparent during stage 1 choices.

Analyses

Simulated learning agents: To check the validity of our custom task and ensure that idealized model-based and model-free learners would show similar patterns of behavior to those with previously used tasks, we simulated 500 model-free and 500 model-based agents "completing" our task. However, we note here that some of the features that made the task unique (trial duration, gamification, spatial layout) were not included in the simulation process.

The simulated model-free learners employed a SARSA algorithm to complete the task (Feher da Silva & Hare, 2020). Initially at $t = 1$ (the first trial), the model-free values of all actions (Q-values) that can be taken at all stages are set to 0.5 (i.e. $Q_1^{MF}(s, a) = 0.5$ for all potential stages and actions). Then, model-free learners update the values of chosen actions. Second-stage action values (at the planets) are updated as the current value, plus the product of the stage two learning rate (α_2) and the reward prediction error at stage 2 (δ_t^2), which is defined as

$$\delta_t^2 = r_t - Q_t^{MF}(s_2, a_2) \quad (1)$$

where r_t is the reward for trial t (either 1 for a coin, or 0 for no coin). Thus, stage 2 Q-values are updated as follows:

$$Q_{t+1}^{MF}(s_2, a_2) = Q_t^{MF}(s_2, a_2) + \alpha_2 \delta_t^2 \quad (2)$$

Then, the Q-value of the chosen first-stage action a_1 at first stage state s_1, the value is also updated according to the products of the stage 1 learning rate (α_1) and reward prediction errors

at *both* stages. Here, the reward prediction error at stage 2 (δ_t^1), was defined as the difference between Q values for the chosen stage 1 and stage 2 actions:

$$\delta_t^1 = Q_t^{MF}(s_2, a_2) + Q_t^{MF}(s_1, a_1) \quad (3)$$

Thus, the Q-value for first stage actions was update as:

$$Q_{t+1}^{MF}(s_1, a_1) = Q_{t+1}^{MF}(s_1, 1) + \alpha_1 \delta_t^1 + \alpha_1 \lambda \delta_t^2 \quad (4)$$

where λ represents the 'eligibility' parameter (which can range from 0-1) governing how well reward prediction errors are back-propagated to previous states and actions. Model-based learners performed updating identically at stage 2 such that $Q_t^{MF}(s_2, a_2) = Q_t^{MB}(s_2, a_2)$. However, model-based agents calculated values for stage 1 actions at the time of decision making using both the values of stage 2 actions and the probabilities that stage 1 actions will transition to stage 2 states as follows:

$$Q_t^{MB}(s_1, a_1) = \sum_{s_2 \in S} P(s_2|s_1, a_1) \, max_{a_2 \in A} Q_t^{MB}(s_2, a_2) \quad (5)$$

where $P(s_1, a_1)$ represents the probability of a given stage 1 action a_1 resulting in a transition to a stage 2 state. S is the set of all possible stage 2 states (both planets), and A is the set of actions available at a given stage 2 state (left or right alien). Thus, action values at stage 1 are calibrated based on the products of the likelihood of transitioning to each stage 2 stage and the maximum action value (best alien) if that stage 2 state is reached.

Both model-free and model-based agents made all decisions with likelihoods governed by a softmax function with inverse temperature parameter β (Daw, 2011), such that higher values of β resulted in heavier weighting of differences in Q-values for any given decision and shifting the balance towards exploitation (whereas lower β increased likelihood of 'exploration'). Both types of agents also were set with a "perseveration" parameter (drawn from a uniform distribution with range [0.01, 0.2]), such that they exhibited a mild tendency to be slightly more likely to repeat previous stage 1 choices than switch. Simulations did not include any such perseveration parameter for stage 2 choices.

Logistic regression models of consecutive trials: For both simulated and human learners in the current study, respectively, we used logistic regression models of consecutive trials to estimate the probability of repeated stage 1 choices as a function of the reward (reward versus no reward) and transition type (common versus rare) on the previous trial. Prior work has demonstrated that ideal model-free agents will more likely repeat stage 1 choices ('stay') if the previous trial resulted in a reward, regardless of whether the previous trial's transition was common or rare (Daw, 2011). Thus, model-free learners will show a main effect of reward, but not a main effect of transition type or interaction. On the other hand, ideal model-based learners will be most likely to repeat stage 1 choices if the previous trial was rewarded and a common transition, or if the previous trial was not rewarded and a rare transition. Thus, purely model-based learners will not show any main effects of reward or transition type, but will show a reward X transition type interaction on probability of staying at stage 1. The R syntax for this model was as follows:

*stay1~last_reward*last_transition + (last_reward*last_transition|id), family = bernoulli(link = "logit")*

In addition to this group-level model, we fit a similar model with a single between-participants term for age, as well as terms for age X reward, age X transition, and age X reward X transition interactions, to ask whether model-free or model-based learning signatures changed as a function of age. The syntax for this model with age terms is as below:

*stay1~last_reward*last_transition*age + (last_reward*last_transition|id), family = bernoulli(link = "logit")*

All logistic regression models were fit using multilevel Bayesian estimation using the brms R package (Bürkner, 2019), with all terms allowed to vary across participants (or simulated agents).

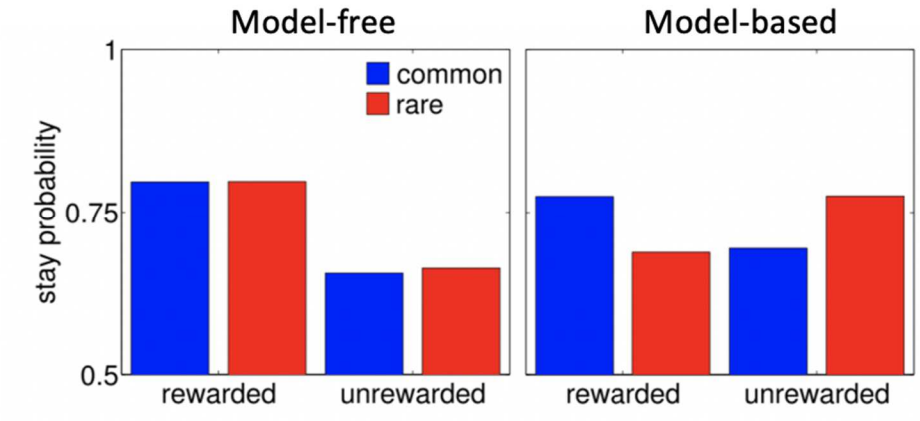


Figure 3.2: Idealized stage 1 decisions of model-free (left) and model-based learners reproduced from Daw et al. (2011). X axis positions represent whether the *previous* trial resulted in a reward or not, and color represents whether the previous trial had a common (blue) or rare (transition). Y axis values represent the likelihood that a learner will "stay", or choose the same stage action as the previous trial.

Knowledge of task structure: To probe participants' understanding of the task structure, we asked them a series of questions after completing the task. Questions asked participants to identify the transition structure most generally (Fig. 3.3, left), as well as which stage 2 state each stage 1 choice most often led to (Fig. 3.3, center & right). For the latter two questions, stage 1 choices (portals) were shown in the center of the bottom of the screen, rather than on the side presented during task trials, such that participants could not rely on spatial information about portals to infer their transition probabilities with stage 2 states.

To assess implicit knowledge of the transition structure, we also measured stage 2 reaction times immediately following common versus rare transitions. Previous work has found participants across ages generally respond more slowly at stage 2 following rare transitions, and that larger rare > common reaction time differences are associated with higher usage of model-based strategies (Decker et al., 2016; Nussenbaum et al., 2020).

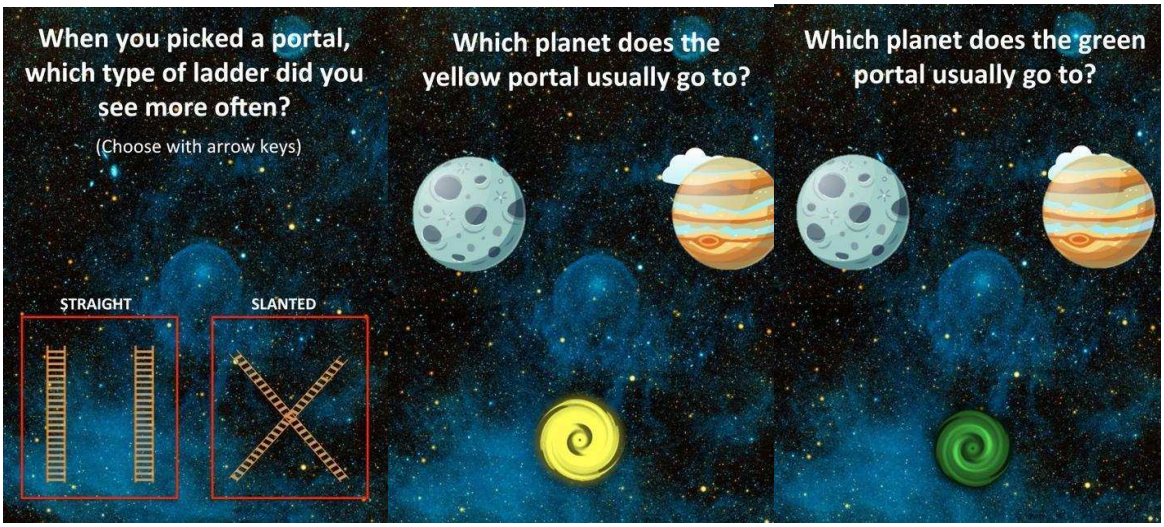


Figure 3.3: Assessment of participants' explicit knowledge of the task structure. For each question, participants could press the left or right arrow key to indicate their answer. Participants answered these questions after completing all trials of the task.

Equivalence of Stimulus, Action, and Spatial Location: Within the current task paradigm, as well as each of the three previously collected datasets, the positions of stimuli at stage 1 (portals in the current paradigm, spaceships in existing datasets) and stage 2 (aliens) were always located on the same sides of the screen. In addition, in the current paradigm, the stage 2 states (planets) and associated actions (aliens) were always visible on screen in the same position (see Fig. 2.1). Thus, unlike previous versions of the task used with adults in which stimuli shifted in their relative positions (Daw et al., 2011), all stage 1 and stage 2 choices were associated with a certain direction (i.e. left or right), and a certain button press (i.e. left arrow key, right arrow key). Because choices were then synonymous with both spatial locations and actions (or motor behaviors), we examined learning behaviors based on spatial and action-based information rather than the stimuli on the screen (Shahar et al., 2019).

Exploratory analyses of stage 2 stay behaviors: After observing that, contrary to our expectations, participants showed neither the canonical signatures for model-free nor model-

based learning at stage 1 (Fig. 2.1, we sought to further characterize participant learning through analyses of decisions at stage 2. In particular, to identify whether such behaviors at stage 2 were specific to our modified paradigm, we conducted secondary analysis of three previously collected datasets using the spaceship version of the two-stage task. Because rewards occur directly after stage 2 decision during the two-stage task, model-based and model-free learners are theorized to behave identically in their decision making at this stage (e.g. because there are no future states for a model-based learner to take into account; Doll et al., 2015). Thus, other than analyses of reaction times at stage 2 following rare versus common decisions as an indicator of awareness of transition structure (Decker et al., 2016), few studies have focused on decision strategies at the second stage of the two-stage task.

Here, in parallel to often-used analyses of stay/switch behaviors at stage 1, we examined the tendency of participants to stay versus switch choices at stage 2. We defined a "stay" at stage 2 as pressing the same button at stage 2 as the previous trial, regardless of state. Under reward learning algorithms, a learning agent would be more likely to stay with the same choice at stage 2 following a reward on the previous trial, provided that the agent is at the same stage 2 state. In fact, previous investigations have used such behaviors as an inclusion criterion under the logic that participants who do not stay with previously rewarded stage 2 choices at the same state are not "pursuing reward" (Decker et al., 2016). However, if the learning agent is at a different stage 2 state compared to the last trial, whether a reward was obtained on that trial should not influence the agent's choice. A reward prediction error on the previous trial will only update the Q-value for the chosen stage 2 action, therefore the relative probability of choices at a different stage 2 state on the subsequent trial will not be different based on whether a reward was obtained. Thus,

neither model-free nor model-based learners, as typically defined, should show effects of previous rewards on stage 2 stay behaviors if at a different stage 2 state compared to the last trial.

To explore stage 2 stay behaviors as a function of reward on the last trial and whether participants were at the same stage 2 state, we fit similar multilevel logistic regression models to those used with stage 1 behavior. For both the simulated agents and for each dataset (PACCT, Decker, Potter, & Nussenbaum) separately, we fit the same model, where *state_match* was coded as a 1 when the participant was at the same stage 2 state as the previous trial, and a 0 when not. To quantify whether stage 2 stays were more likely following stage 1 stays, models also included an interaction term for stage 1 stays (*stay1*) on the current trial as follows:

*stay2 ~ state_match*last_reward*stay1 + (state_match*last_reward*stay1 | id),*
family = bernoulli(link = 'logit')

To examine whether any such behaviors showed age-related differences, we also fit the above model with an added term for age, including interactions of age with all parameters. The model was fit separately to each dataset as follows:

*stay2 ~ state_match*last_reward*stay1*age + (state_match*last_reward*stay1 | id), family = bernoulli(link = 'logit')*

Because such higher-order interactions tend to be much lower powered and prone to errors of estimation than do lower-order interactions or main effects (Gelman, 2018), we examined whether such interaction effects generalized across all 4 datasets.

Stage 2 reaction times: In efforts to understand whether participants made faster decisions when executing reward-contingent action sequences, we fit a multilevel linear regression model to participants' stage 2 reaction times. We first within-participant z-scored stage 2 reaction times such that each participant's mean reaction time was set to 0 and the standard deviation set to 1. We then modeled these z-scored reaction times as a function of reward on the previous trial (*last_reward),* stays/switches at both stage 1 (*stay1*) and stage 2 (*stay2*) of the current trial, as well as the transition type (*transition)* on the current trial:

*rt2_z ~ stay1*stay2*last_reward*transition + (stay1*stay2*last_reward*transition | id)*

We examined the relative speed of reward-contingent action sequences through the parametrization of a 3-way *stay1 x stay2 x last_reward* interaction on stage 2 reaction times. Additionally, we examined whether such effects differed when the current trial transition was common versus rare through the 4-way interaction of *stay1 x stay2 x last_reward x transition.*

Model-fitting*:* For all analyses, Bayesian multilevel regression models were fit using Markov Chain Monte Carlo estimation using the brms package in R (Bürkner, 2019). Unless otherwise indicated, 4 chains of 2000 iterations (1000 warmup) were run for each model. All models used package-default weakly informative priors unless otherwise indicated. 95% posterior intervals (PI) and posterior predictive intervals of expected values are reported using the quantile method. All data analyses were done using R, and visualizations were created using the ggplot2 package (Wickham et al., 2019).

3.3 Results

Simulated learning agents within the current task paradigm: As a check of our task parameters, we simulated decisions for 500 model-free and 500 model-based learning agents. We then fit multilevel logistic regression models to the model-free and model-based agents separately to stage 1 stay/switch behaviors contingent on the last trial reward and transition type. As previously demonstrated with two-stage paradigms (Daw et al., 2011), model-free learners most often stayed with stage 1 choices (made the same choice as the previous trial) following rewarded trials, but did not show responsiveness to the transition type of the previous trial (Fig. 3.4A). Model-based learners, on the other hand, showed an interaction effect between the reward and transition type on the last trial. (Fig. 3.4B) Model-based learners were most likely to stay with stage 1 choices following common rewarded trials and rare unrewarded trials, and relatively less likely to stay following rare rewarded trials and common unrewarded trials. As previously demonstrated, model-free and model-based learners performed identically following trials with a common transition, but showed different patterns of stage 1 stay behaviors following a rare transition.

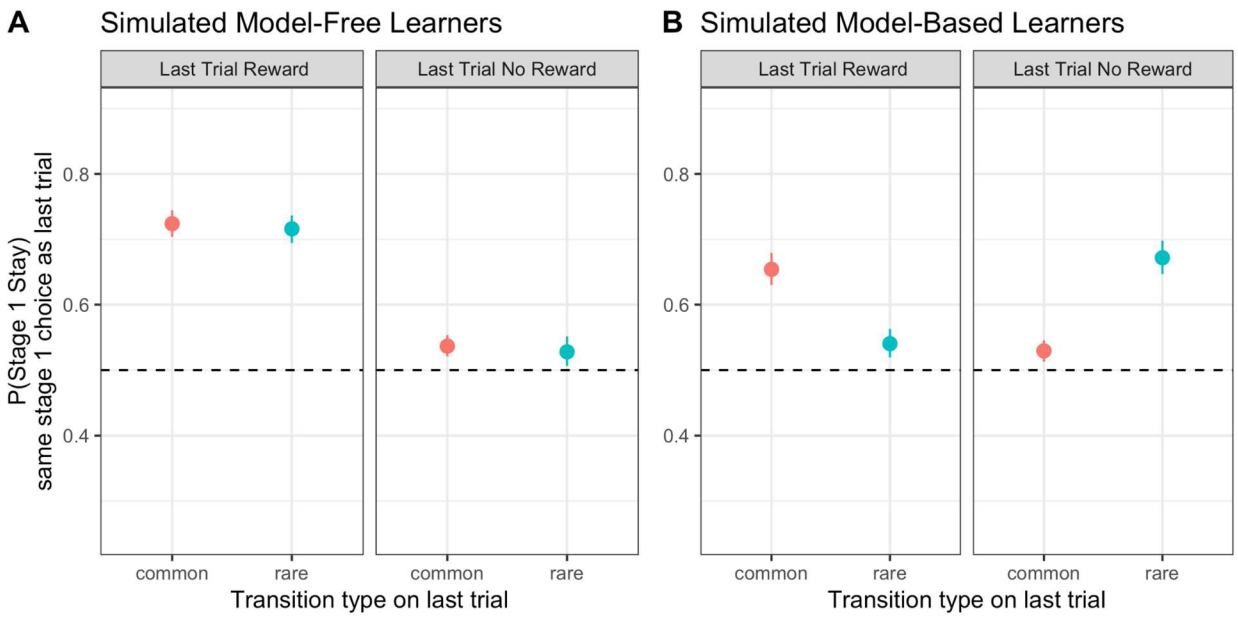


Figure 3.4: Simulated model-free (A) and model-based (B) learning agents' stay/switch behaviors at the first stage of the current task paradigm. Panels indicate whether the last trial resulted in reward (left) or not (right), while the x-axis indicates common verus rare transitions on the previous trial, and the y-axis indicates the probability of staying with the stage 1 choice made on the last trial. The model-free learners (left) and model-based learners (right) both showed typical behaviors, such that model-free learners were more likely to stay at stage 1 following rewarded trials with both rare and common transitions, whereas model-based learners were more likely to stay at stage 1 following rewarded common trials and non-rewarded rare trials.

Neither model-free nor model based learning signatures in the current study: We examined participants' use of model-free and model-based learning strategies in our modified paradigm using logistic regression analyses of stage 1 stay/switch behaviors. Contrary to our expectations, participants showed neither signatures typical of model-free nor model-based reward learning, as we found little evidence for a main effect of reward on the previous trial, or previous reward X previous trial transition interactions (see Fig. 3.5 & Table 3.2). However, we found a main effect of previous trial transition type, such that participants were more likely to stay at stage 1 following a rare trial regardless of whether a reward was obtained. Within the

current cohort ages 7-13 years, we found no age-related differences in any of the stage 1 behaviors analyzed (see Table 3).

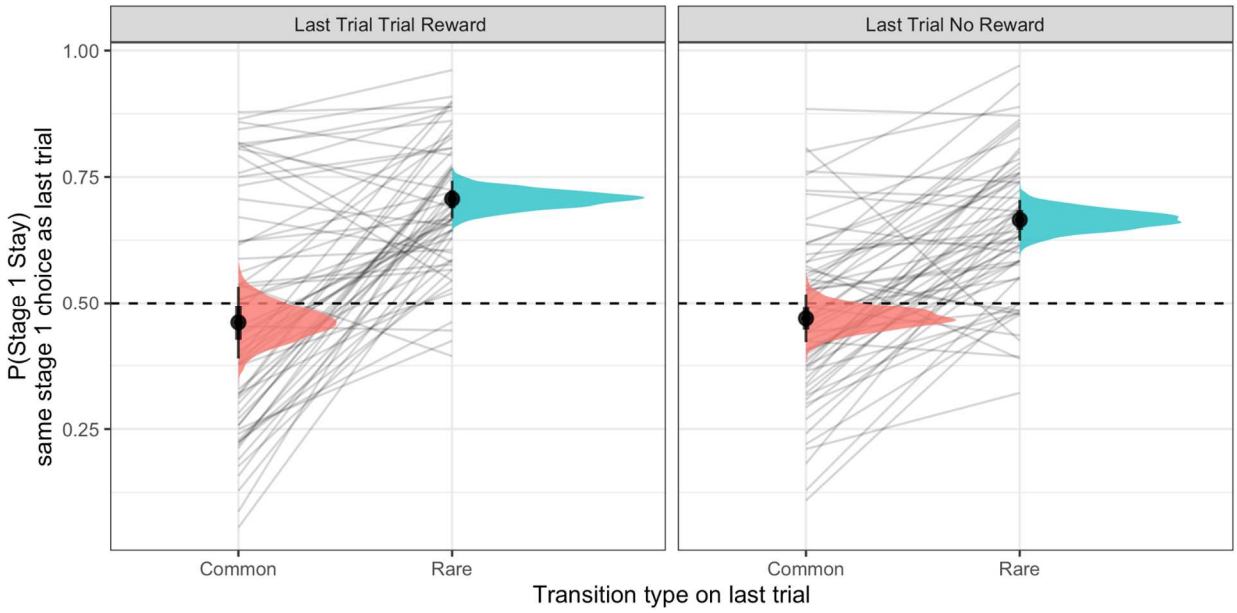


Figure 3.5: Stage 1 stay/switch behaviors in the current paradigm as a function of reward and transition type on the previous trial. Panels indicate whether the participant received a reward (left) or not (right) on the previous trial, and color indicates whether the previous trial was a common (red) or rare (blue) transition. The y-axis indicates the probability of stage 1 stays (i.e. making the same choice as the previous trial) under each condition. Shaded distributions indicate expected values of the posterior predictive distribution for each condition from the multilevel logistic regression model. Thick and thin error bars represent 80% and 95% posterior intervals, respectively. Thin gray lines represent raw proportions of stays under each condition for each participant (1 line is 1 participant).

Predictor	*Mean Estimate*	*95% Posterior Interval*

Intercept	-0.12	[-0.324, 0.06]
Last Trial Reward	-0.03	[-0.266, 0.196]
Last Trial Rare	0.81	[0.621, 1.002]
Last Trial Reward X Last Trial Rare	0.23	[-0.092, 0.525]

Table 3.2: Logistic regression parameters for effects of previous reward and previous transition on subsequent stage 1 stay/switch choices in the modified two-stage paradigm. Parameter estimates are in log odds.

Predictor	*Mean Estimate*	*95% Posterior Interval*
Intercept	0.32	[-0.754 , 1.421]
Last Trial Reward	0.10	[-1.287 , 1.397]
Last Trial Rare	0.44	[-0.732 , 1.56]
Age	-0.05	[-0.161 , 0.065]
Last Trial Reward X Last Trial Rare	0.03	[-1.673 , 1.807]
Last Trial Reward X Age	-0.01	[-0.147 , 0.126]
Last Trial Rare X Age	0.04	[-0.076 , 0.158]
Last Trial Reward X Last Trial Rare X Age	0.02	[-0.161 , 0.196]

Table 3.3: Logistic regression parameters for age-related differences in the modified two-stage paradigm in effects of previous reward and previous transition on subsequent stage 1 stay/switch choices. Age (years, range = [7,13]) is treated continuously, and parameter estimates are in log odds.

Knowledge of transition structure in the current study: We assessed participants' understanding of the task structure in the current study using both explicit questions at the end of the task and reaction time measures. A majority of participants (64.5%, 71%, and 72.6%) correctly answered each of 3 binary-choice questions on the task structure after completing the study (Fig. 3.6A). Although a sizable minority of participants did not correctly answer some questions on the task structure, responses overall nevertheless suggested that most participants understood the task transition structure. In addition, a multilevel linear regression indicated that participants were slower on average (Δ_{RT}=141.5ms, 95% CI [116.6, 165.5]) to make stage 2 choices following rare transitions compared to common transitions during the current trial (Fig. 3.6B), indicating behavioral sensitivity to transition type.

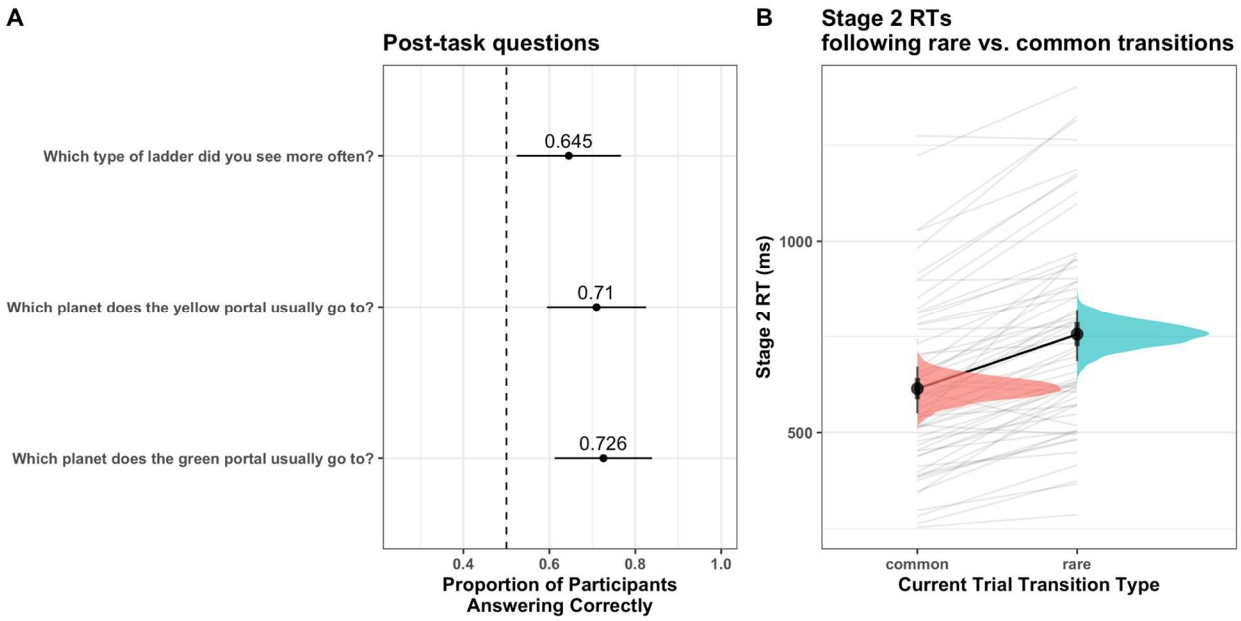


Figure 3.6: A: Participant responses to questions on explicit knowledge of transition structure. Error bars indicate bootstrapped 95% confidence intervals. **B:** Participants' stage 2 reaction times as a function of transition type and trial. Distributions indicate posterior predictive distributions for the mean stage 2 reaction time, and thick and thin error bars represent 80% and 95% posterior intervals, respectively. Gray lines represent individual participants' average reaction times (1 line = 1 participant). Stage 2 reaction times were consistently slower following

rare transitions compared to common transitions throughout the duration of the task, suggesting that participants were responsive to the transition structure.

Reward-contingent stage 2 behaviors common across datasets: After initial analyses indicated unexpected behavioral patterns at the first stage of the current paradigm (see Fig. 3.5), we worked to more fully characterize participants' decision-making strategies. To this end, we conducted parallel analyses of stage 2 behaviors in the current paradigm as well as three previously collected two-stage datasets (see Table 1). We first examined stage 2 stay versus switch behaviors, where stays were defined as an identical button press at stage 2 compared to the last trial (regardless of stage 2 state). Simulated model-free (β = 1.69, 95% PI [1.40, 1.98]) and model-based (β = 1.54, 95% PI [1.30, 1.78]) learners both showed an interaction effect between stage 2 location and last trial rewards on stage 2 staying behavior. Learners under both algorithms were more likely to stay at stage 2 following rewards when at the same stage 2 state (i.e. planet) as the previous trial, but showed no reward-contingent staying when at the other stage 2 state compared to the last trial (Fig. 3.7B). Such patterns did not differ based on stage 1 stay versus switch behaviors (e.g. stage 2 location X last trial reward X stage 1 stay interaction; $\beta_{\text{model-free}}$ = -0.17, 95% PI [-0.44, 0.09]; $\beta_{\text{model-based}}$ = 0.08, 95% PI [-0.19, 0.34]). Thus, these simulations demonstrated that under model-free and model-based algorithms, previous reward-contingent choices at stage 2 were specific to when learners were at the same stage 2 state as the last trial.

Like the simulated model-free and model-based learners, participants in each cohort demonstrated interaction effects between stage 2 location and last trial rewards on stage 2 staying behavior (see Fig. 3.7; β_{PACCT} = 0.69, 95% PI [0.38, 1.00]; β_{Decker} = 1.06, 95% PI [0.68, 1.47], $\beta_{\text{Nussenbaum}}$ = 1.43, 95% PI [1.09, 1.79]; β_{Potter} = 1.65, 95% PI [1.14, 2.16]). However, unlike the

simulated learners, participants in each cohort showed main effects of last trial rewards on stage 2 staying behavior, such that last trial rewards increased the probability of stage 2 stays even on trials where participants were at the other stage 2 state (i.e. planet) compared to the last trial (Fig. 3.7A). In addition, unlike simulated learners, participants in each cohort showed last reward X stage 1 stay interaction ($\beta_{\text{PACCT}} = 0.60$, 95% PI [0.39, 0.81]; $\beta_{\text{Decker}} = 0.50$, 95% PI [0.28, 0.72], $\beta_{\text{Nussenbaum}} = 0.24$, 95% PI [0.07, 0.41]; $\beta_{\text{Potter}} = 0.37$, 95% PI [0.11, 0.63]), such that they were more likely to stay at stage 2 following staying at stage 1 on the same trial (Fig. 3.8).

While the above patterns of stage 2 behavior were consistent across all datasets, within the current paradigm (PACCT) patterns of reward-contingent staying behavior were weaker on trials where the current stage 2 state matched the last trial. In particular, participants were no more likely to stay at stage 2 following a reward when the stage 2 state matched the last trial compared to when it did not match (see Fig. 3.7A; $\Delta_{\text{stay\_probability}} = 0.01$, 95% PI [-0.03, 0.06]).

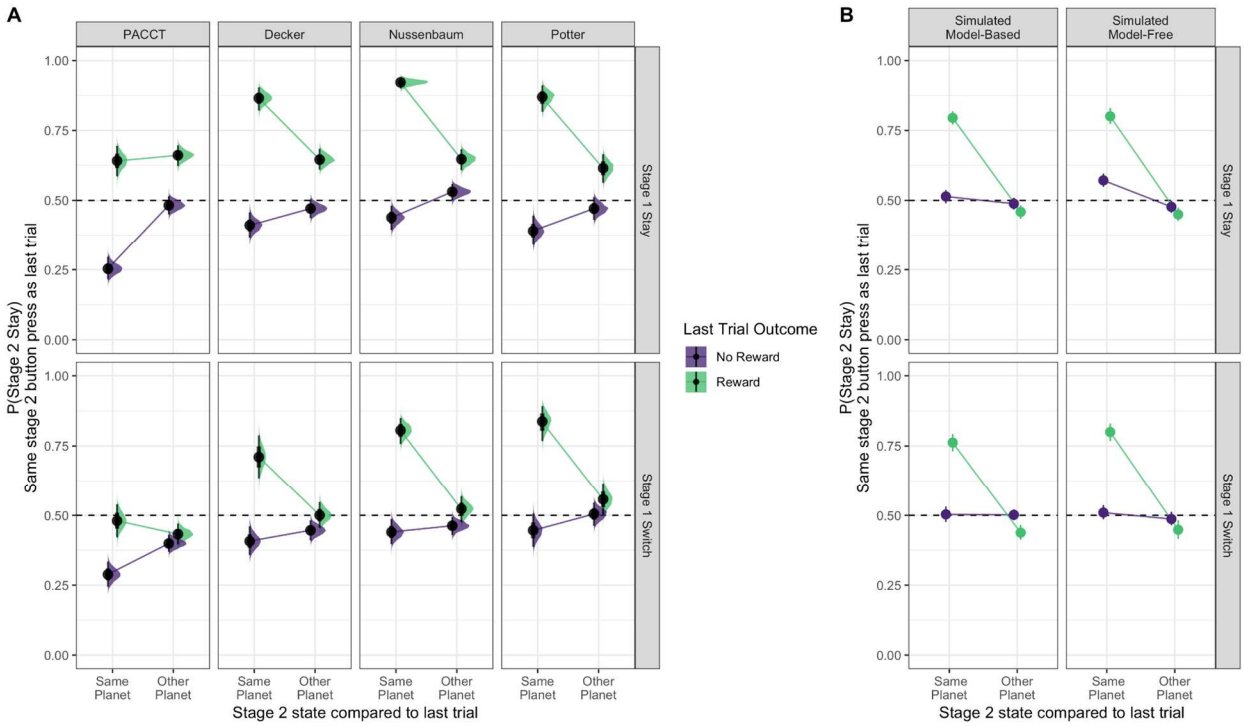


Figure 3.7: Reward-contingent stay/switch behaviors at stage 2. Plots show the probability of stage 2 stays (defined as the same stage 2 button press as the previous trial) on the y-axis. Stage 2 stay probability is shown as a function of reward on the last trial (purple = no reward, green = reward), whether the stage 2 state is the same (same planet) or different (other planet) compared to the last trial, and whether the participant stayed with the same stage 1 choice as the previous trial (top row = stage 1 stay, bottom row = stage 1 switch). **A**: Plots for participants in the PACCT, Decker, Nussenbaum, and Potter datasets. Distributions indicate expected values of the posterior predictive distribution for each condition from the multilevel logistic regression model. Thick and thin error bars represent 80% and 95% posterior intervals, respectively. **B:** Simulated model-based (left) and model-free (right) learners completing the current study paradigm. For all conditions, chance performance = 50%.

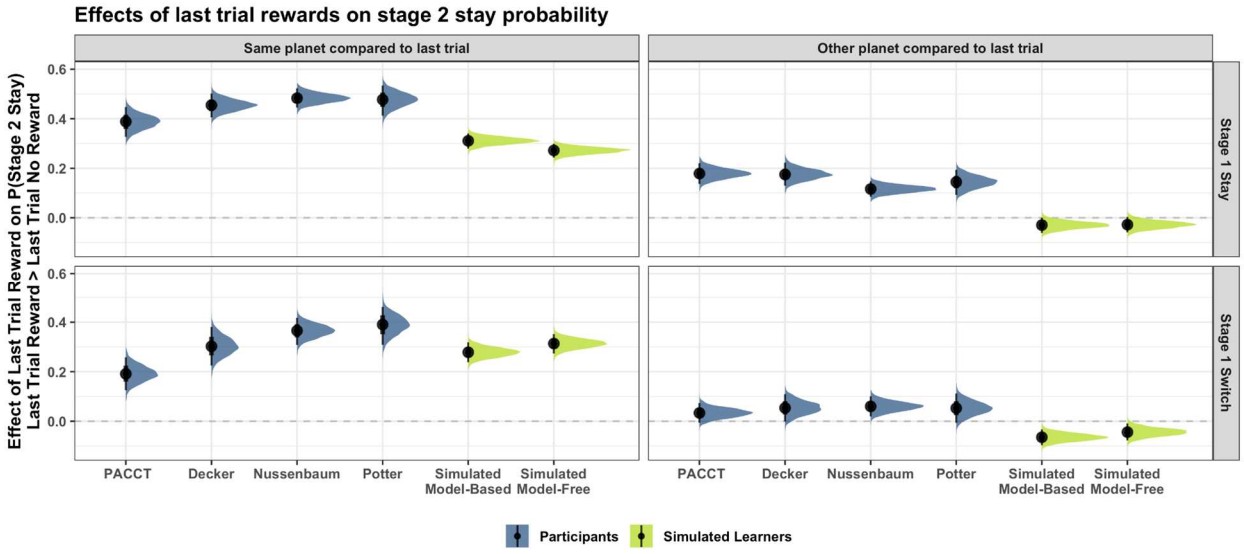


Figure 3.8: Parametrization of reward-contingent stay/switch behaviors at stage 2. Each posterior distribution shows the contrast of the probability of stage 2 stays when the last trial was rewarded > last trial not rewarded. 0 indicates no effect, such that stage 2 stays were equally likely following rewarded > non-rewarded trials. Posterior contrasts such for reward effects are subset based on whether the stage 2 state (planet) is the same as the previous trial (columns) and whether or not the participant stayed at stage 1 (rows). Posterior distributions for such contrasts are shown for each cohort (blue) and both model-free and model-based simulated learners (yellow). Thick and thin error bars represent 80% and 95% posterior intervals, respectively.

Faster reaction times for reward-contingent stage 2 behaviors across datasets: Given evidence across cohorts for reward-contingent staying behaviors at stage 2, especially following stage 1 stays, we investigated differences in stage 2 reaction times as a function of stays at both stages, last trial rewards, and current trial transition type using multilevel linear regression models fit to each cohort. Negative 3-way stage 1 X stage 2 X last trial reward interaction terms (Fig. 3.9B, top panel) for each cohort indicated that following rewards, participants made stage 2 choices faster on average if staying at both stages of the current trial (Fig. 3.9A). We did not find strong evidence for a 4-way stage 1 X stage 2 X last trial reward X current trial transition interaction across cohorts (Fig. 3.9B, bottom panel), such that such reward-contingent speeded reaction times when staying at both stages occurred whether the current trial contained a rare or common transition (although stage 2 reaction times were slower in general following rare transitions, see Fig. 3.6B & Nussenbaum et al., 2020).

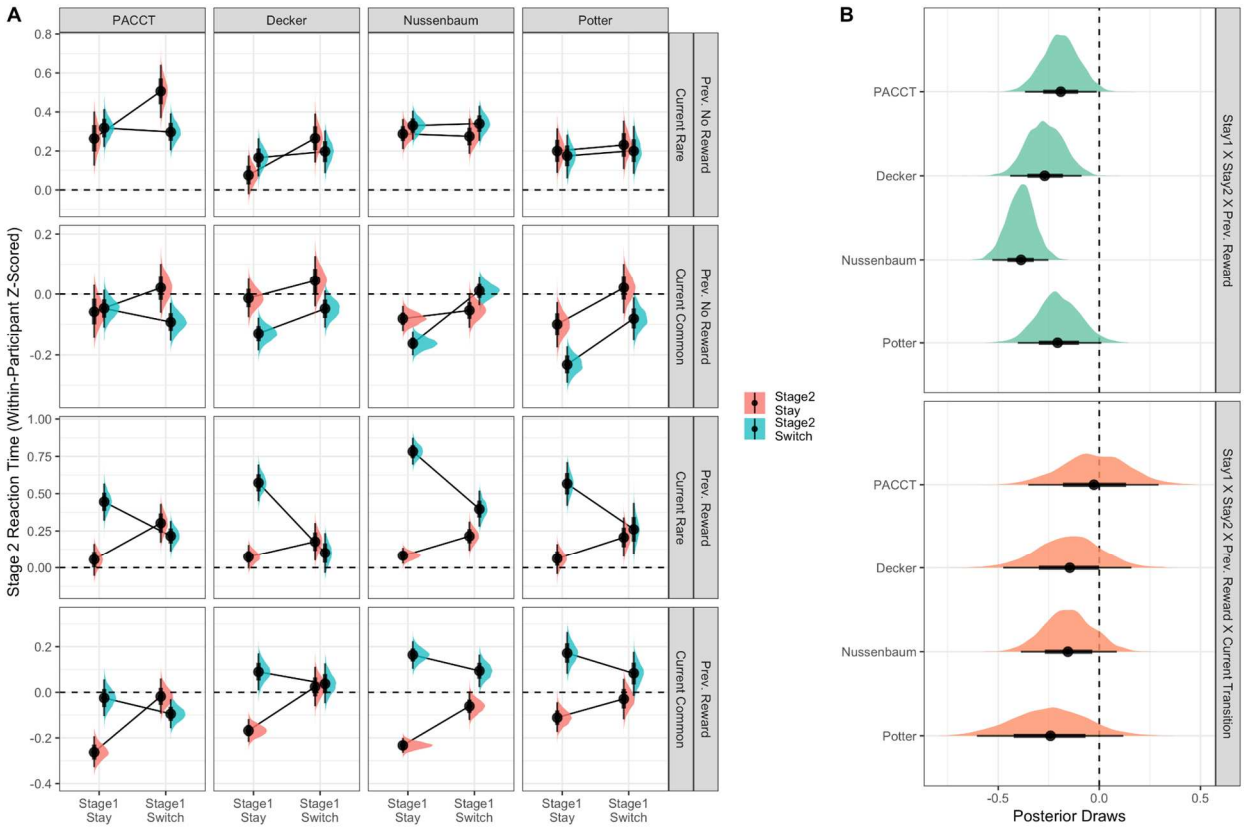


Figure 3.9: Reaction times contingent on rewards and stay/switch behaviors at stage 2. **A**: Plots show within-participant z-scored stage 2 reaction times on the y-axis as a function of reward on the last trial, whether participants stayed at both stage 1 (x-axis) and stage 2 (red=stay, blue=switch), and whether the current trial is rare or common (rows). Plots for participants in the PACCT, Decker, Nussenbaum, and Potter datasets. Distributions indicate expected values of the posterior predictive distribution for each condition from the multilevel linear regression model. **B:** Posterior distributions for each dataset for key model parameters from the regression model visualized in A. The top panel represents 3-way stay1 X stay2 X last trial reward interactions, where negative estimates indicate stronger evidence of reward-contingent speeding of stage 2 choices on trials with stay decisions at both stages. The bottom panel indicates 4-way stay1 X stay2 X last trial reward X current transition interactions, where more negative estimates indicate stronger reward-contingent speeding on trials with stays at both stages for rare trials, compared to common trials. Thick and thin error bars represent 80% and 95% posterior intervals, respectively.

Age-related differences in reward-contingent behavioral sequences: We examined age-related differences in each cohort in reward-contingent behaviors at stage 2. With age treated as a continuous between-participants variable in multilevel regression models, we explored age-

related differences in the reward contingency of stage 2 stays on trials where participants stayed at stage 1 (Fig. 3.10). In all datasets other than the current paradigm (PACCT), age was positively associated with the probability of staying at stage 2 following a previously rewarded trial if at the same planet again (Fig. 3.10A). No consistent age-related differences in reward-contingent stage 2 stay behaviors were observed on trials where the stage 2 location was the other planet compared to the previous trial (Fig. 3.10B). However, across the entire age range studied, participants were more likely to stay at stage 2 following rewards (in Fig. 3.10A green ribbons always above purple), consistent with prior work indicating age-related increase in inverse temperature (i.e. lower "decision noise"; Eckstein et al., 2021; Nussenbaum & Hartley, 2019).

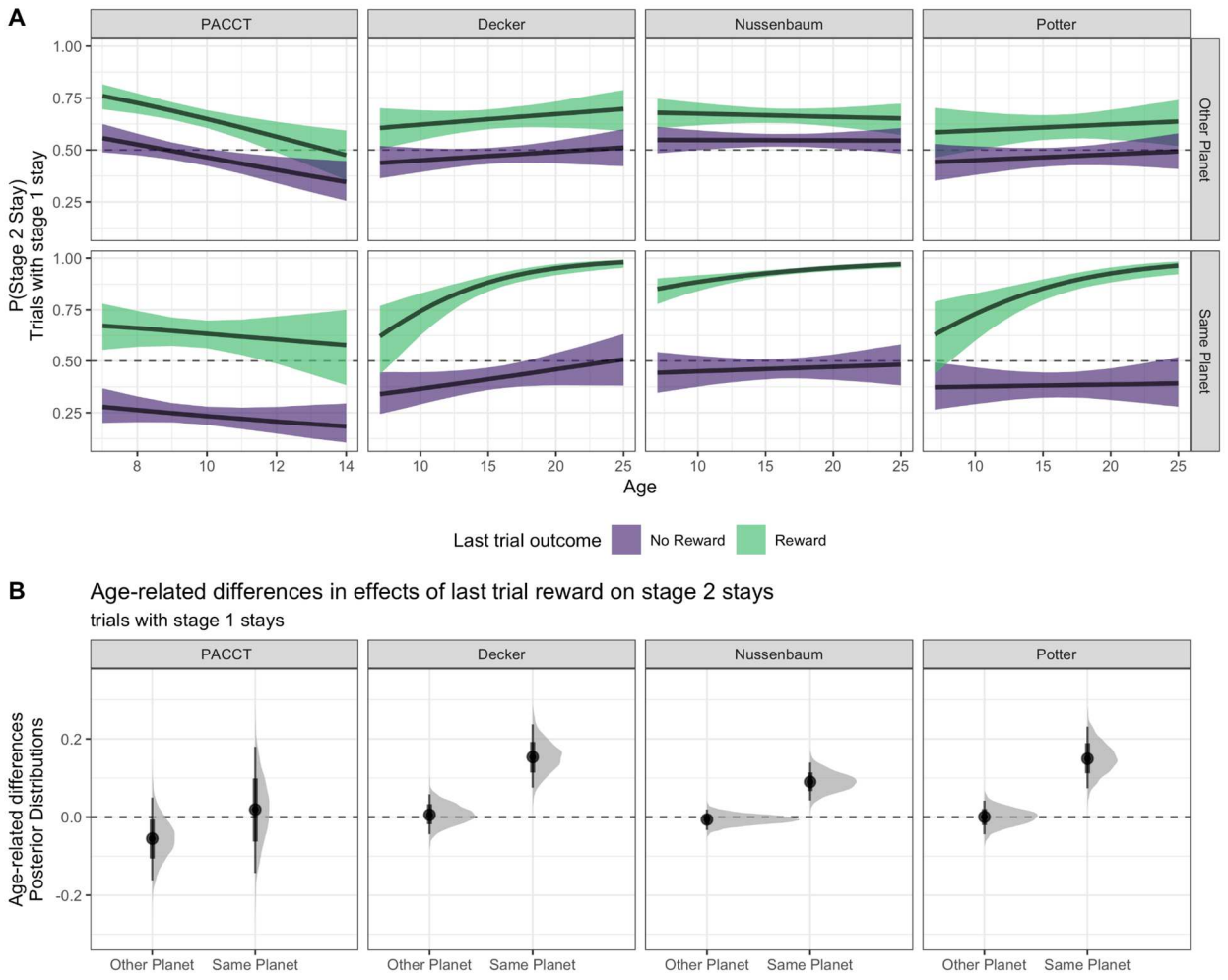


Figure 3.10: Age-related differences in reward-contingent stage 2 stay behaviors on trials with stage 1 stays. **A**: Model posterior predictions for the mean probability of stage 2 stays as a function of age, stage 2 location (other planet vs. same planet), and last trial reward (green) versus no reward (purple). Lines show predicted means and shaded areas represent 95% PIs. We note that the x-axis range is different (7-14) for the PACCT dataset given the narrower age range of this cohort. **B:** Posterior distributions for age X last trial reward interactions on stage 2 stays on trials with stage 1 stays. Distributions are shown for each cohort, and separately for subsets of trials where participants were at the same stage 2 location (same planet) versus other location (other planet) compared to the previous trial. Positive values indicate positive associations between age and effects of reward on stage 2 stays. Thick and thin error bars represent 80% and 95% posterior intervals, respectively.

3.4 Discussion

The current study aimed initially to adapt a child-friendly two-stage paradigm to encourage model-based learning. However, contrary to what was predicted, youth aged 7-13 displayed little evidence of typical model-free or model-based signatures in completing this paradigm despite evidence of understanding the task transition structure. Several factors in the current study may have contributed to behavioral patterns quite different from previously reported, including task 'gamification', shortened instructions and practice trials, the spatial layout of the transition structure, shortened trial durations, long study sessions, and lack of monetary incentives for task performance. Follow-up explorations of stage 2 behaviors in data collected under this modified paradigm, as well as three previously collected developmental datasets using the two-stage task, indicated that participants across cohorts used reward-contingent strategies not explained by either typical model-free or model-based algorithms. In particular, speeded reward-contingent motor sequences indicated that participants may tend to use spatial or motor cues to generate action policies during reward learning even when not instructed to. More broadly, the current work highlights both the fact that model-free and model-based behaviors may be highly paradigm-specific, as well as that even within the two-stage paradigm, youth may pursue strategies "beyond" those expected by either algorithm. We discuss each of these points in greater detail below.

Neither typical model-free nor model-based strategies under the current paradigm: Contrary to our expectations, when completing our modified two-stage paradigm, participants demonstrated neither the signatures of model-free nor model-based learning typically observed in studies using the two-stage task (Daw et al., 2011; Decker et al., 2016; Nussenbaum et al., 2020; Otto, Raio, et al., 2013; Potter et al., 2017; Sharp et al., 2016). Participants showed neither

main effects on stage 1 staying behaviors of last trial rewards (indicative of model-free learning) nor reward X transition interactions (indicative of model-based learning). Instead, at stage 1 participants showed only sensitivity to rare versus common transitions on the previous trial, such that they stayed at stage 1 more often following rare transitions (see Fig. 3.5). Such unexpected behavior was not likely due to a lack of awareness of the task structure or a lack of attention to the task, as participants' choices at both stages and reaction times at stage 2 reflected a sensitivity to the transition structure (Figs 3.5-6).

Potential mechanisms for stage 1 stays following rare transitions on the last trial: Unlike prior studies, participants repeat stage 1 choices more often after rare trials, regardless of reward (Fig. 3.5). One possibility is that such behaviors represent "novelty-seeking" or "exploration" in the sense that rare transitions are novel relative to common ones (Gopnik, 2020). While repeating a stage 1 choice following a rare transition does not increase the chance of a subsequent rare transition or increase the likelihood of obtaining novel information per se, participants could have found rare trials more salient due to their relative novelty (Galván, 2010; Lloyd et al., 2021). Participants may have believed that rare transitions marked stage 1 choices that would reveal new information (Krebs et al., 2009). Participants also may have believed erroneously that rare transitions were associated with rewards, particularly because only common transitions were shown in the instructions and practice trials in the current study.

On the other hand, participants may have repeated stage 1 choices following rare trials in efforts to reach the stage 2 state (planet) not visited on the previous trial. When asked to describe their decision-making strategies, several participants indicated that they tried to "switch planets each time," which could be accomplished more often under a policy of switching after common trials and staying after rare trials at stage 1. Such a policy is also consistent with a "depletion

model" where stage 2 states or actions are believed to be less likely to yield rewards on multiple consecutive visits. Indeed, some participants reported beliefs such as "if an alien gave a reward, it probably didn't have one next time so I tried to go to a different planet". Such statements are consistent with a model-based policy under such a depletion model, as they indicate prospective decision-making to avoid previously rewarding stage 2 actions. Finally, it is also possible that participants found rare transitions more intrinsically rewarding than the "coins" they received as rewards, especially given that no true monetary compensation was provided to extrinsically motivate better task performance. Overall, mechanisms for repeated stage 1 choices following rare trials most likely stem from interpretations of the task structure not aligned with reward learning algorithms, as such behavior is not beneficial towards earning rewards.

An additional consideration is that participants' task behaviors did not always match the strategies they explicitly self-reported. One reason for this is that younger participants may not have been able to explicitly describe the algorithms by which they made decisions, despite readily using them. Additionally, this observation is supported by prior work indicating that participants may use both declarative and non-declarative "habitual" processes in learning from rewards (Foerde et al., 2006; Otto, Gershman, et al., 2013). Thus, participants' descriptions of their own strategies, while not incorrect per se, may have been incomplete. On the other hand, because explicit questions on task strategies were at the end of the task, participants may have reported using strategies that they used towards the latter trials. Especially if participants shifted behaviors over the course of the task, such explicit responses may not represent behavioral patterns from earlier trials.

Potential contributors to differences with prior developmental two-stage task studies*:* As previous work has demonstrated within adults (Feher da Silva & Hare, 2020),

unclear or misleading task instructions may have caused participants' altered interpretations of the current paradigm and resultant patterns of behavior. One potentially influential factor is that in the current study, instructions were shortened compared to previous versions of the spaceship two-stage task (Daw et al., 2011; Decker et al., 2016; Nussenbaum et al., 2020; Potter et al., 2017), and the number of practice trials was reduced from 20-50 to 2. Especially given that previous studies have given participants more extensive practice (with different stimuli, 20 or 50 trials) and specific instructions on the transition structure, such scaffolding may have narrowed the space of strategies participants chose from (i.e. towards model-free or model-based algorithms) to complete the task. In contrast, participants may have begun the current task paradigm with fewer priors guiding their decision-making processes, and therefore turned to alternative action policies.

In addition, the spatial layout of the task transition matrix on the screen may have encouraged alternative decision-making policies. Previous versions of the two-stage task have not imparted specific spatial context to the transition matrix (i.e. stage 2 stimuli replace stage 1 stimuli on screen). While we intended for the representation of common transitions with vertical ladders and rare transitions with diagonal ones to encourage model-based strategies under the logic that a model with spatial properties would be easier to use, it is possible that this oriented participants away from such reward learning strategies at the first stage. Because adolescent and adult participants have previously been shown to use spatial-motor cues in lieu of model-based strategies (Shahar et al., 2019), the mapping of transition structures with spatial relationships may have encouraged such reward-independent strategies.

The "gamification" of the task may also have contributed to the unusual stage 1 behaviors observed under the current paradigm. While the "space treasure" theme and music, sounds, and

animations were added to boost engagement with the task, these changes may also have distracted participants or oriented their focus towards reward-independent information. Further, because durations of transitions, rewards, and inter-trial intervals were shortened relative to previous versions of the task, the quickened pace of the task may have encouraged participants to pursue less cognitively burdensome strategies. While model-based learning has been suggested to particularly make demands on working memory (Kool et al., 2016; Otto, Raio, et al., 2013; Potter et al., 2017), both model-free and model-based learners at minimum maintain expected value computations for stage 2 actions. Fatigue that participants may have experienced as they completed this task during approximately 4-hour study visits could also have contributed towards participants' selection of less working memory-intensive strategies. While any combination of the above factors may contribute to the unusual patterns of stage 1 choices under the current paradigm, future work directly isolating and manipulating such factors will be needed to better understand their influence on such behaviors.

Reward-Contingent Spatial-Motor Behaviors: After initial analyses indicated that participants under the current paradigm did not show typical signatures of model-free or model-based learning, we set out to characterize other factors explaining their behavior and investigate whether these behaviors were common to three other developmental cohorts (Decker et al., 2016; Nussenbaum et al., 2020; Potter et al., 2017). Particularly because stimuli were associated with distinct spatial-motor information at both stages of the task across datasets, we investigated whether participants displayed reward-contingent behaviors based on such mappings even when such choices would not increase reward probability. Consistent with prior work (Shahar et al., 2019), participants repeated stage 2 choices that had been rewarded on the previous trial, even when at a different state compared to the last trial (Figs 3.7-8). During these trials where the

current stage 2 state did not match that of the last trial, simulated model-free and model-based agents did not show reward-contingent stage 2 stay behaviors. On such trials, repeated stage 2 choices indicate a binding of spatial (left vs. right) or motor (a particular button press) information with reward, despite the fact that such information is irrelevant to reward probability. Participants were also more likely to repeat stage 2 actions after having repeated their stage 1 action compared to the previous trial, even in such reward-independent situations, indicating erroneous association of spatial-motor information with reward at both stages of the task. Such reward-independent decision-making was not associated with age (Fig. 3.10).

Across all datasets, we also found evidence that participants chose actions more quickly when behaving consistently with such reward-contingent spatial-motor mappings. Following rewarded trials, participants' stage 2 reaction times were fastest on trials where they repeated both stage 1 and stage 2 choices (Fig. 3.9). The same pattern was not true following trials without rewards. Thus, participants may have chosen to quickly repeat "sequences" of button presses yielding rewards on the previous trial. As previously suggested (Shahar et al., 2019), these behaviors suggest "embodied" decision-making strategies as learners attempt to decide between visual stimuli and their own actions in assigning credit following rewards (McDougle et al., 2016).

Potential mechanisms underlying mapping of rewards to spatial-motor behavior: While credit assignment to reward-independent spatial and motor information will not increase rewards (Shahar et al., 2019), spatial and motor cues are not purely reward-independent in the studied datasets. As particular button presses are associated with actions at both stages, action policies assigning value only to motor execution at each stage (irrespective of the transition structure) will maximize reward likelihood when the current stage 2 state is the same as the last

one. Thus, participants pursuing such a policy may have identified an adaptive "shortcut" for completing the task. Even though this strategy fails on trials with incongruent consecutive stage 2 states, repeating sequences of button presses following a rewarded trial may be a form of model-free computation with low working memory demands.

Alternatively, it is possible that participants may not be mapping rewards to spatial-motor behaviors, but rather assuming non-independence of the expected values of stage 2 actions. For example, participants may update the expected values of non-chosen (or non-visited) states and actions based on rewards (Biderman & Shohamy, 2021). Because independently drifting reward probabilities of each of the four stage 2 choices may be particularly difficult for participants to update and maintain in working memory (Master et al., 2020), consolidating the values of stage 2 choices may be an adaptive step.

Limitations: The current findings are limited by several factors that may be addressed in future work. First, because only youth ages 7-13 participated in the current paradigm (PACCT), direct comparisons of this cohort with older participants in the previously collected datasets was not possible. That no participants in the current study completed both the spaceship version of the two-stage task (Decker et al., 2016) and our modified version also precludes making strong conclusions about which factors may have contributed to differences between task paradigms. In addition, the current study only examined learning behaviors as a function of the previous trial, rather than using multi-step approaches (Miller et al., 2016) or full reinforcement learning models (Daw, 2011; Eckstein et al., 2021). Future work using such methods may be able to characterize model-free, model-based, and spatial-motor decision-making strategies in more depth.

Conclusion

The current investigation found evidence for both study-specific and more general patterns of behavior among developmental cohorts beyond the model-free and model-based learning strategies often studied in the context of the two-stage sequential decision-making tasks. Differential patterns of behavior within a modified version of the two-stage paradigm highlighted the potential sensitivity of participants' learning strategies to factors including task instructions, trial duration, study visit duration, and incentivization of rewards (Smid et al., 2020). Consistent with recent work (Shahar et al., 2019), participants ages 7-25 demonstrated decision-making based on spatial-motor information even in contexts where such information was reward-independent. While the use of such spatial-motor cues occurred in addition to, rather than instead of, model-free and model-based strategies, the present findings add evidence to recent calls to characterize reward learning across a broader set of potential policies (Daw, 2018; Feher da Silva & Hare, 2020; Momennejad et al., 2017). In particular, while previous work has examined associations between model-based control and stress (Otto, Raio, et al., 2013), working memory (Potter et al., 2017; Silva et al., 2018), depression (Heller et al., 2018), and dopaminergic function (Doll et al., 2016; Sharp et al., 2016), future work may benefit from characterizing similar associations using spatial-motor decision-making strategies.

Conclusion

The chapters in this dissertation demonstrate the use of multiverse approaches within several research contexts, in particular in examining robustness of results to researcher decisions, optimizing such decisions, and exploring multiple psychological mechanisms. Yet, the core components of multiverses are far from new concepts. Indeed, much prior work has developed and used statistical analyses of sensitivity (Saltelli & Annoni, 2010; Vincent et al., 2008) and model comparison (Hastie et al., 2001). However, multiverse approaches provide a principled set of approaches for increasing the breadth of decision points examined and quantifying their impacts (Niso et al., 2022). Multiverses also offer a systematic way of computing data reliability metrics necessary for many statistical paradigms across many decision points. Multiverse analyses can also function as organized checks of internal validity through probing possible combinations of confounding variables (Frank, 2000), or convergent validity through systematic comparisons with a different measurement of the same construct (for example, fMRI signal could be compared with another brain imaging modality). Further, specification curves in particular allow for visualization of research findings that emphasize the totality of the evidence across multiple reasonable strategies.

One key consideration for multiverse approaches is that due to the sheer number of analyses, multiverses by definition involve multiple comparisons (Steegen et al., 2016). Thus, in classic multiverse analyses where the goal is to determine where there exists a robust finding across specifications, individual specifications should not be overinterpreted. Rather, joint inference can be made based on a parameter summarizing all analyses, such as the median estimate across specifications (Simonsohn et al., 2020). Crucially, researchers should report all tested decision points and interpret equally reasonable analyses with equal evidential weight to

avoid selective reporting based on the desirability of results (i.e. p-hacking; Liu et al., 2020; Wicherts et al., 2016). Multiverse decision points can also be preregistered to maximize confirmatory value and minimize risk of "p-hacking" (Flournoy et al., 2020; Olsson-Collentine et al., 2021). When multiple comparisons between individual specifications is central to a research question, both preregistration and replication of multiverse findings in independent datasets can help build evidential strength.

A further practical issue is that in theory, specification curve analyses should test "all reasonable choices" by definition (Simonsohn et al., 2015). While researchers may then, justifiably, want to consider *all* possible combinations of decisions when analyzing their data or optimizing a method, such comprehensiveness is often computationally intractable. Particularly for computationally demanding analyses (such as fMRI preprocessing), multiverses may be most feasible when targeting a small number of decision points believed to be most consequential. Further, the computational burden of larger multiverse analyses raises issues of accessibility of research computing resources (Chalker et al., 2020) as well as environmental concerns (because high computing can require vast energy consumption; Anthony et al., 2020; Lannelongue et al., 2021). While there are extant resources for lowering such computational burdens (Fan et al., 2022; Lawrence et al., 2021) and widening the accessibility of performance computing (Pestilli, 2022; Towns et al., 2014), limiting multiverse size and scope is nevertheless crucial in many contexts.

Although limiting the expansion of multiverse analyses can be difficult, researchers can use several strategies to limit their scope. First, researchers may be able to eliminate decision points if alternatives are not truly "arbitrary" or if one choice is unambiguously superior to another based on prior empirical or theoretical work (Simonsohn et al., 2020; Steegen et al.,

2016). In contrast to "full multiverses" with all possible choices included, researchers can also use psychometric properties and theoretical domain knowledge to construct smaller "principled multiverse-style analyses" best suited for a question of interest (Del Giudice & Gangestad, 2021). More cumulatively, multiverse investigations can seek to optimize decision-making for studied choices so that such decisions will no longer be arbitrary for future investigations. Thus, multiverse studies can eliminate decision points for subsequent ones.

When multiverses reveal that different, or even seemingly arbitrary, choices yield varied results, researchers may experience discouragement or lack of confidence in their work. On one hand, lack of confidence in a particular result is warranted given high sensitivity to reasonable analytic choices (Botvinik-Nezer et al., 2020; Orben & Przybylski, 2019; Steegen et al., 2016). A self-correcting science requires that researchers alter methodologies given strong evidence that previously used techniques are flawed (Vazire & Holcombe, 2021). Yet, at least some variation in results across analysis strategies would be expected in any circumstance, even when a finding is highly "robust" (Patel et al., 2015). Sensitivity to a particular choice may not doom a measurement or analysis strategy, but can instead indicate that alternative choices truly answer distinct empirical questions. Further, researchers with multiverse results indicating high sensitivity to analytical decisions might be encouraged by the fact that such findings are often methodological contributions alone. That a particular methodological decision contributes great variance to results might enhance understanding of the method, or prompt fruitful examination into the research questions in which the method might be most aptly used.

The chapters in this dissertation emphasize the utility of multiverse analyses across several different contexts and types of research questions. Although much of the work here focuses on functional MRI, multiverse strategies can be broadly valuable in any studies where

researchers face multiple reasonable options. Indeed, recent multiverse studies have addressed topics including electrophysiology (Clayson et al., 2021; Gordillo et al., 2022), the microbiome (Nearing et al., 2022), experience sampling methods (Weermeijer et al., 2022), and clinical subtyping (Beijers et al., 2020). Because multiverses and specification curves are still relatively young as formal statistical methods, they may hold much promise for future methodological development and applied use across many fields of research.

References

Abdulrahman, H., & Henson, R. N. (2016). Effect of trial-to-trial variability on optimal event-related fMRI design: Implications for Beta-series correlation and multi-voxel pattern analysis. *NeuroImage*, *125*, 756–766. https://doi.org/10.1016/j.neuroimage.2015.11.009

Achenbach, T. M. (1991). *Integrative Guide for the 1991 CBCL/4-18, Ysr, and Trf Profiles* (1st US-1st Printing edition). Univ Vermont/Dept Psychiatry.

Achterberg, M., & van der Meulen, M. (2019). Genetic and environmental influences on MRI scan quantity and quality. *Developmental Cognitive Neuroscience*, *38*, 100667. https://doi.org/10.1016/j.dcn.2019.100667

Adolphs, R. (2008). Fear, Faces, and the Human Amygdala. *Current Opinion in Neurobiology*, *18*(2), 166–172. https://doi.org/10.1016/j.conb.2008.06.006

Allen, J. L., Lavallee, K. L., Herren, C., Ruhe, K., & Schneider, S. (2010). DSM-IV criteria for childhood separation anxiety disorder: Informant, age, and sex differences. *Journal of Anxiety Disorders*, *24*(8), 946–952. https://doi.org/10.1016/j.janxdis.2010.06.022

Anthony, L. F. W., Kanding, B., & Selvan, R. (2020). Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. *ArXiv:2007.03051 [Cs, Eess, Stat]*. http://arxiv.org/abs/2007.03051

Baird, A. A., Gruber, S. A., Fein, D. A., Mass, L. C., Steingard, R. J., Renshaw, P. F., Cohen, B. M., & Yurgelun-todd, D. A. (1999). Functional Magnetic Resonance Imaging of Facial Affect Recognition in Children and Adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, *38*(2), 195–199. https://doi.org/10.1097/00004583-199902000-00019

Bassett, D. S., Wymbs, N. F., Porter, M. A., Mucha, P. J., Carlson, J. M., & Grafton, S. T. (2011). Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences*, *108*(18), 7641–7646. https://doi.org/10.1073/pnas.1018985108

Bates, D., & Bolker, M. M. and B. (2011). *lme4: Linear mixed-effects models using S4 classes* (0.999375-39) [Computer software]. http://www.idg.pl/mirrors/CRAN/web/packages/lme4/

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., & Fox, J. (2019). *lme4: Linear Mixed-Effects Models using "Eigen" and S4* (1.1-21) [Computer software]. https://CRAN.R-project.org/package=lme4

Beeley, C. (2013). *Web Application Development with R using Shiny*. Packt Publishing Ltd.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A Component Based Noise Correction Method (CompCor) for BOLD and Perfusion Based fMRI. *NeuroImage*, *37*(1), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Beijers, L., Loo, H. M. van, Romeijn, J.-W., Lamers, F., Schoevers, R. A., & Wardenaar, K. J. (2020). Investigating data-driven biological subtypes of psychiatric disorders using specification-curve analysis. *Psychological Medicine*, 1–12. https://doi.org/10.1017/S0033291720002846

Belloy, M. E., Naeyaert, M., Abbas, A., Shah, D., Vanreusel, V., van Audekerke, J., Keilholz, S. D., Keliris, G. A., Van der Linden, A., & Verhoye, M. (2018). Dynamic resting state fMRI analysis in mice reveals a set of Quasi-Periodic Patterns and illustrates their

relationship with the global signal. *NeuroImage*, *180*(Pt B), 463–484. https://doi.org/10.1016/j.neuroimage.2018.01.075

Beretta, L., & Santaniello, A. (2016). Nearest Neighbor Imputation Algorithms: A Critical Evaluation. *BMC Medical Informatics and Decision Making*, *16*(Suppl 3). https://doi.org/10.1186/s12911-016-0318-z

Biderman, N., & Shohamy, D. (2021). Memory and decision making interact to shape the value of unchosen options. *Nature Communications*, *12*(1), 4648. https://doi.org/10.1038/s41467-021-24907-x

Birmaher, B., Brent, D. A., Chiappetta, L., Bridge, J., Monga, S., & Baugher, M. (1999). Psychometric Properties of the Screen for Child Anxiety Related Emotional Disorders (SCARED): A Replication Study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *38*(10), 1230–1236. https://doi.org/10.1097/00004583-199910000-00011

Birn, R. M., Cornejo, M. D., Molloy, E. K., Patriat, R., Meier, T. B., Kirk, G. R., Nair, V. A., Meyerand, M. E., & Prabhakaran, V. (2014a). The Influence of Physiological Noise Correction on Test–Retest Reliability of Resting-State Functional Connectivity. *Brain Connectivity*, *4*(7), 511–522. https://doi.org/10.1089/brain.2014.0284

Birn, R. M., Cornejo, M. D., Molloy, E. K., Patriat, R., Meier, T. B., Kirk, G. R., Nair, V. A., Meyerand, M. E., & Prabhakaran, V. (2014b). The Influence of Physiological Noise Correction on Test–Retest Reliability of Resting-State Functional Connectivity. *Brain Connectivity*, *4*(7), 511–522. https://doi.org/10.1089/brain.2014.0284

Birn, R. M., Diamond, J. B., Smith, M. A., & Bandettini, P. A. (2006). Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. *NeuroImage*, *31*(4), 1536–1548. https://doi.org/10.1016/j.neuroimage.2006.02.048

Birn, R. M., Smith, M. A., Jones, T. B., & Bandettini, P. A. (2008). The Respiration Response Function: The temporal dynamics of fMRI signal fluctuations related to changes in respiration. *NeuroImage*, *40*(2), 644–654. https://doi.org/10.1016/j.neuroimage.2007.11.059

Bloom, P. A., VanTieghem, M., Gabard-Durnam, L., Gee, D. G., Flannery, J., Caldera, C., Goff, B., Telzer, E. H., Humphreys, K. L., Fareri, D. S., Shapiro, M., Algharazi, S., Bolger, N., Aly, M., & Tottenham, N. (2022). Age-related change in task-evoked amygdala—prefrontal circuitry: A multiverse approach with an accelerated longitudinal cohort aged 4–22 years. *Human Brain Mapping*, *n/a*(n/a). https://doi.org/10.1002/hbm.25847

Bossier, H., Roels, S. P., Seurinck, R., Banaschewski, T., Barker, G. J., Bokde, A. L. W., Quinlan, E. B., Desrivières, S., Flor, H., Grigis, A., Garavan, H., Gowland, P., Heinz, A., Ittermann, B., Martinot, J.-L., Artiges, E., Nees, F., Orfanos, D. P., Poustka, L., … Moerkerke, B. (2020). The empirical replicability of task-based fMRI as a function of sample size. *NeuroImage*, *212*, 116601. https://doi.org/10.1016/j.neuroimage.2020.116601

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Boubela, R. N., Kalcher, K., Huf, W., Seidel, E.-M., Derntl, B., Pezawas, L., Našel, C., & Moser, E. (2015). FMRI measurements of amygdala activation are confounded by stimulus correlated signal fluctuation in nearby veins draining distant brain regions. *Scientific Reports*, *5*(1), 10499. https://doi.org/10.1038/srep10499

Bowring, A., Maumet, C., & Nichols, T. E. (2019). Exploring the impact of analysis software on task fMRI results. *Human Brain Mapping*, *40*(11), 3362–3384. https://doi.org/10.1002/hbm.24603

Bridgeford, E. W., Wang, S., Yang, Z., Wang, Z., Xu, T., Craddock, C., Dey, J., Kiar, G., Gray-Roncal, W., Priebe, C. E., Caffo, B., Milham, M., Zuo, X.-N., Reproduciblity, C. for R. and, & Vogelstein, J. T. (2020). Big Data Reproducibility: Applications in Brain Imaging. *BioRxiv*, 802629. https://doi.org/10.1101/802629

Brosch, J. R., Talavage, T. M., Ulmer, J. L., & Nyenhuis, J. A. (2002). Simulation of human respiration in fMRI with a mechanical model. *IEEE Transactions on Biomedical Engineering*, *49*(7), 700–707. https://doi.org/10.1109/TBME.2002.1010854

Bryce, N. V., Flournoy, J. C., Guassi Moreira, J. F., Rosen, M. L., Sambook, K. A., Mair, P., & McLaughlin, K. A. (2021). Brain parcellation selection: An overlooked decision point with meaningful effects on individual differences in resting-state functional connectivity. *NeuroImage*, *243*, 118487. https://doi.org/10.1016/j.neuroimage.2021.118487

Bürkner, P.-C. (2017). *brms: Bayesian Regression Models using Stan* (1.6.1) [Computer software]. https://cran.r-project.org/web/packages/brms/index.html

Bürkner, P.-C. (2019). *brms: Bayesian Regression Models using "Stan"* (2.10.0) [Computer software]. https://CRAN.R-project.org/package=brms

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Caballero-Gaudes, C., & Reynolds, R. C. (2017). Methods for cleaning the BOLD fMRI signal. *NeuroImage*, *154*, 128–149. https://doi.org/10.1016/j.neuroimage.2016.12.018

Calvo, M. G., & Lundqvist, D. (2008). Facial expressions of emotion (KDEF): Identification under different display-duration conditions. *Behavior Research Methods*, *40*(1), 109–115. https://doi.org/10.3758/BRM.40.1.109

Carpenter, K. L. H., Angold, A., Chen, N.-K., Copeland, W. E., Gaur, P., Pelphrey, K., Song, A. W., & Egger, H. L. (2015). Preschool Anxiety Disorders Predict Different Patterns of Amygdala-Prefrontal Connectivity at School-Age. *PLOS ONE*, *10*(1), e0116854. https://doi.org/10.1371/journal.pone.0116854

Casey, B. J., Galván, A., & Somerville, L. H. (2016). Beyond simple models of adolescence to an integrated circuit-based account: A commentary. *Developmental Cognitive Neuroscience*, *17*, 128–130. https://doi.org/10.1016/j.dcn.2015.12.006

Chalker, A., Hillegas, C. W., Sill, A., Broude Geva, S., & Stewart, C. A. (2020). Cloud and on-premises data center usage, expenditures, and approaches to return on investment: A survey of academic research computing organizations. In *Practice and Experience in Advanced Research Computing* (pp. 26–33). Association for Computing Machinery. https://doi.org/10.1145/3311790.3396642

Chang, C., & Glover, G. H. (2009). Relationship between respiration, end-tidal CO2, and BOLD signals in resting-state fMRI. *NeuroImage*, *47*(4), 1381–1393. https://doi.org/10.1016/j.neuroimage.2009.04.048

Charlton, P. H., Birrenkott, D. A., Bonnici, T., Pimentel, M. A. F., Johnson, A. E. W., Alastruey, J., Tarassenko, L., Watkinson, P. J., Beale, R., & Clifton, D. A. (2018). Breathing Rate Estimation From the Electrocardiogram and Photoplethysmogram: A Review. *IEEE Reviews in Biomedical Engineering*, *11*, 2–20. https://doi.org/10.1109/RBME.2017.2763681

Chen, G., Pine, D. S., Brotman, M. A., Smith, A. R., Cox, R. W., Taylor, P. A., & Haller, S. P. (2022). Hyperbolic trade-off: The importance of balancing trial and subject sample sizes in neuroimaging. *NeuroImage*, *247*, 118786. https://doi.org/10.1016/j.neuroimage.2021.118786

Cho, J. W., Korchmaros, A., Vogelstein, J. T., Milham, M., & Xu, T. (2020). Impact of Concatenating fMRI Data on Reliability for Functional Connectomics. *BioRxiv*, 2020.05.06.081679. https://doi.org/10.1101/2020.05.06.081679

Chorpita, B. F., Yim, L., Moffitt, C., Umemoto, L. A., & Francis, S. E. (2000). Assessment of Symptoms of DSM-IV Anxiety and Depression in Children: A Revised Child Anxiety and Depression Scale. *Behaviour Research and Therapy*, *38*(8), 835–855. https://doi.org/10.1016/S0005-7967(99)00130-8

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*(2), 127–137.

Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G. L., Ruparel, K., Shinohara, R. T., Elliott, M. A., Eickhoff, S. B., Davatzikos, C., Gur, R. C., Gur, R. E., Bassett, D. S., & Satterthwaite, T. D. (2017a). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage*, *154*, 174–187. https://doi.org/10.1016/j.neuroimage.2017.03.020

Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G., Ruparel, K., Shinohara, R. T., Elliott, M. A., Eickhoff, S. B., Davatzikos, C., Gur, R. C., Gur, R. E., Bassett, D. S., & Satterthwaite, T. D. (2017b). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage*, *154*, 174–187. https://doi.org/10.1016/j.neuroimage.2017.03.020

Cisler, J. M., Bush, K., & Steele, J. S. (2014). A Comparison of Statistical Methods for Detecting Context-Modulated Functional Connectivity in fMRI. *NeuroImage*, *84*, 1042–1052. https://doi.org/10.1016/j.neuroimage.2013.09.018

Clayson, P. E., Baldwin, S. A., Rocha, H. A., & Larson, M. J. (2021). The data-processing multiverse of event-related potentials (ERPs): A roadmap for the optimization and standardization of ERP processing and reduction pipelines. *NeuroImage*, *245*, 118712. https://doi.org/10.1016/j.neuroimage.2021.118712

Collins, A. G. E., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, *21*(10), 576–586. https://doi.org/10.1038/s41583-020-0355-6

Cosme, D., & Lopez, R. (2020). *Neural indicators of food cue reactivity, regulation, and valuation and their associations with body composition and daily eating behavior*. PsyArXiv. https://doi.org/10.31234/osf.io/23mu5

Cox, R. W. (1996). AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Computers and Biomedical Research, an International Journal*, *29*(3), 162–173.

Craddock, C., Sikka, S., Cheung, B., Khanuja, R., Ghosh, S. S., Yan, C.-G., Li, Q., Lurie, D., Vogelstein, J., Burns, R., Colcombe, S., Mennes, M., Kelly, C., Di Martino, A., Castellanos, F. X., & Milham, M. (2013). *Towards Automated Analysis of Connectomes: The Configurable Pipeline for the Analysis of Connectomes (C-PAC)*. Neuroinformatics, Stockholm, Sweden. https://doi.org/10.3389/conf.fninf.2013.09.00042

Crone, E. A., & Elzinga, B. M. (2015). Changing brains: How longitudinal functional magnetic resonance imaging studies can inform us about cognitive and social-affective growth trajectories. *WIREs Cognitive Science*, *6*(1), 53–63. https://doi.org/10.1002/wcs.1327

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve Frequently Asked Questions About Growth Curve Modeling. *Journal of Cognition and Development*, *11*(2), 121–136. https://doi.org/10.1080/15248371003699969

Dafflon, J., Costa, P. F. D., Váša, F., Monti, R. P., Bzdok, D., Hellyer, P. J., Turkheimer, F., Smallwood, J., Jones, E., & Leech, R. (2020). *Neuroimaging: Into the Multiverse* (p. 2020.10.29.359778). https://doi.org/10.1101/2020.10.29.359778

Daw, N. D. (2011). *Trial-by-trial data analysis using computational models: (Tutorial Review)*. Oxford University Press. http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199600434.001.0001/acprof-9780199600434-chapter-001

Daw, N. D. (2018). Are we of two minds? *Nature Neuroscience*, *21*(11), 1497–1499. https://doi.org/10.1038/s41593-018-0258-2

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The Contribution of Emotion and Cognition to Moral Sensitivity: A Neurodevelopmental Study. *Cerebral Cortex*, *22*(1), 209–220. https://doi.org/10.1093/cercor/bhr111

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From Creatures of Habit to Goal-Directed Learners: Tracking the Developmental Emergence of Model-Based Reinforcement Learning. *Psychological Science*, *27*(6), 848–858. https://doi.org/10.1177/0956797616639301

Del Giudice, M., & Gangestad, S. W. (2021). A Traveler's Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920954925. https://doi.org/10.1177/2515245920954925

Delgado, M. R., Nearing, K. I., LeDoux, J. E., & Phelps, E. A. (2008). Neural Circuitry Underlying the Regulation of Conditioned Fear and Its Relation to Extinction. *Neuron*, *59*(5), 829–838. https://doi.org/10.1016/j.neuron.2008.06.029

Di, X., & Biswal, B. B. (2017). Psychophysiological Interactions in a Visual Checkerboard Task: Reproducibility, Reliability, and the Effects of Deconvolution. *Frontiers in Neuroscience*, *11*. https://doi.org/10.3389/fnins.2017.00573

Di, X., Reynolds, R. C., & Biswal, B. B. (2017). Imperfect (de)Convolution May Introduce Spurious Psychophysiological Interactions and How to Avoid It. *Human Brain Mapping*, *38*(4), 1723–1740. https://doi.org/10.1002/hbm.23413

Di, X., Zhang, Z., & Biswal, B. B. (2020). Understanding psychophysiological interaction and its relations to beta series correlation. *BioRxiv*, 322073. https://doi.org/10.1101/322073

Doll, B. B., Bath, K. G., Daw, N. D., & Frank, M. J. (2016). Variability in Dopamine Genes Dissociates Model-Based and Model-Free Reinforcement Learning. *Journal of Neuroscience*, *36*(4), 1211–1222. https://doi.org/10.1523/JNEUROSCI.1901-15.2016

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience*, *18*(5), 767–772. https://doi.org/10.1038/nn.3981

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*(6), 1075–1081. https://doi.org/10.1016/j.conb.2012.08.003

Dragicevic, P., Jansen, Y., Sarma, A., Kay, M., & Chevalier, F. (2019). Increasing the Transparency of Research Papers with Explorable Multiverse Analyses. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3290605.3300295

Dreyfuss, M., Caudle, K., Drysdale, A. T., Johnston, N. E., Cohen, A. O., Somerville, L. H., Galván, A., Tottenham, N., Hare, T. A., & Casey, B. J. (2014). Teens Impulsively React rather than Retreat from Threat. *Developmental Neuroscience*, *36*(3–4), 220–227. https://doi.org/10.1159/000357755

Duncan, K., Doll, B. B., Daw, N. D., & Shohamy, D. (2018). More Than the Sum of Its Parts: A Role for the Hippocampus in Configural Reinforcement Learning. *Neuron*, *98*(3), 645-657.e6. https://doi.org/10.1016/j.neuron.2018.03.042

Durand, E., Moortele, P.-F. van de, Pachot-Clouard, M., & Bihan, D. L. (2001). Artifact due to B0 fluctuations in fMRI: Correction using the k-space central line. *Magnetic Resonance in Medicine*, *46*(1), 198–201. https://doi.org/10.1002/mrm.1177

Eckstein, M., Wilbrecht, L., & Collins, A. (2021). *What do RL Models Measure? Interpreting Model Parameters in Cognition and Neuroscience*. PsyArXiv. https://doi.org/10.31234/osf.io/e7kwx

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, *1*(1), 56–75. https://doi.org/10.1007/BF01115465

Elliott, M. L., Knodt, A. R., & Hariri, A. R. (2021). Striving toward translation: Strategies for reliable fMRI measurement. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2021.05.008

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis: *Psychological Science*. https://doi.org/10.1177/0956797620916786

Erwin, R. J., Gur, R. C., Gur, R. E., Skolnick, B., Mawhinney-Hee, M., & Smailis, J. (1992). Facial emotion discrimination: I. Task construction and behavioral findings in normal subjects. *Psychiatry Research*, *42*(3), 231–240. https://doi.org/10.1016/0165-1781(92)90115-j

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline

for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Fair, D. A., Miranda-Dominguez, O., Snyder, A. Z., Perrone, A., Earl, E. A., Van, A. N., Koller, J. M., Feczko, E., Tisdall, M. D., van der Kouwe, A., Klein, R. L., Mirro, A. E., Hampton, J. M., Adeyemo, B., Laumann, T. O., Gratton, C., Greene, D. J., Schlaggar, B. L., Hagler, D. J., … Dosenbach, N. U. F. (2020). Correction of respiratory artifacts in MRI head motion estimates. *NeuroImage*, *208*, 116400. https://doi.org/10.1016/j.neuroimage.2019.116400

Falahpour, M., Refai, H., & Bodurka, J. (2013). Subject specific BOLD fMRI respiratory and cardiac response functions obtained from global signal. *NeuroImage*, *72*, 252–264. https://doi.org/10.1016/j.neuroimage.2013.01.050

Fan, C. C., Palmer, C. E., Iversen, J. R., Pecheva, D., Holland, D., Frei, O., Thompson, W. K., Hagler, D. J., Andreassen, O. A., Jernigan, T. L., Nichols, T. E., & Dale, A. M. (2022). *FEMA: Fast and efficient mixed-effects algorithm for population-scale whole-brain imaging data* (p. 2021.10.27.466202). bioRxiv. https://doi.org/10.1101/2021.10.27.466202

Feher da Silva, C., & Hare, T. A. (2020). Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour*, *4*(10), 1053–1066. https://doi.org/10.1038/s41562-020-0905-y

Fields, A., Bloom, P. A., VanTieghem, M., Harmon, C., Choy, T., Camacho, N. L., Gibson, L., Umbach, R., Heleniak, C., & Tottenham, N. (2021). Adaptation in the face of adversity: Decrements and enhancements in children's cognitive control behavior following early

caregiving instability. *Developmental Science*, *n/a*(n/a). https://doi.org/10.1111/desc.13133

Finn, E. S., & Rosenberg, M. D. (2021). Beyond fingerprinting: Choosing predictive connectomes over reliable connectomes. *NeuroImage*, *239*, 118254. https://doi.org/10.1016/j.neuroimage.2021.118254

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., & Constable, R. T. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, *18*(11), 1664–1671. https://doi.org/10.1038/nn.4135

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Flournoy, J. C., Vijayakumar, N., Cheng, T. W., Cosme, D., Flannery, J. E., & Pfeifer, J. H. (2020). Improving practices and inferences in developmental cognitive neuroscience. *Developmental Cognitive Neuroscience*, *45*, 100807. https://doi.org/10.1016/j.dcn.2020.100807

Foerde, K., Knowlton, B. J., & Poldrack, R. A. (2006). Modulation of competing memory systems by distraction. *Proceedings of the National Academy of Sciences*, *103*(31), 11778–11783. https://doi.org/10.1073/pnas.0602659103

Forbes, E. E., Phillips, M. L., Silk, J. S., Ryan, N. D., & Dahl, R. E. (2011). Neural Systems of Threat Processing in Adolescents: Role of Pubertal Maturation and Relation to Measures of Negative Affect. *Developmental Neuropsychology*, *36*(4), 429–452. https://doi.org/10.1080/87565641.2010.550178

Fox, M. D., Zhang, D., Snyder, A. Z., & Raichle, M. E. (2009). The Global Signal and Observed Anticorrelated Resting State Brain Networks. *Journal of Neurophysiology*, *101*(6), 3270–3283. https://doi.org/10.1152/jn.90777.2008

Francis, G., Last, C. G., & Strauss, C. C. (1987). Expression of separation anxiety disorder: The roles of age and gender. *Child Psychiatry and Human Development*, *18*(2), 82–89. https://doi.org/10.1007/BF00709952

Frank, K. A. (2000). Impact of a Confounding Variable on a Regression Coefficient. *Sociological Methods & Research*, *29*(2), 147–194. https://doi.org/10.1177/0049124100029002001

Frank, M. J., & Badre, D. (2012). Mechanisms of Hierarchical Reinforcement Learning in Corticostriatal Circuits 1: Computational Analysis. *Cerebral Cortex*, *22*(3), 509–526. https://doi.org/10.1093/cercor/bhr114

Frankenhuis, W. E., Panchanathan, K., & Barto, A. G. (2019). Enriching behavioral ecology with reinforcement learning methods. *Behavioural Processes*, *161*, 94–100. https://doi.org/10.1016/j.beproc.2018.01.008

Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., & Turner, R. (1996). Movement-related effects in fMRI time-series. *Magnetic Resonance in Medicine*, *35*(3), 346–355. https://doi.org/10.1002/mrm.1910350312

Gabry, J., Ali, I., Brilleman, S., Novik, J. B., AstraZeneca, Wood, S., Development, R. C., Bates, D., Maechler, M., Bolker, B., Walker, S., Burkner, P.-C., Ripley, B., Venables, W., & Goodrich, B. (2019). *Rstanarm: Bayesian Applied Regression Modeling via Stan* (2.19.2) [Computer software]. https://CRAN.R-project.org/package=rstanarm

Galván, A. (2010). Adolescent development of the reward system. *Frontiers in Human Neuroscience*, *4*. https://www.frontiersin.org/article/10.3389/neuro.09.006.2010

Gee, D. G., Humphreys, K. L., Flannery, J., Goff, B., Telzer, E. H., Shapiro, M., Hare, T. A., Bookheimer, S. Y., & Tottenham, N. (2013). A Developmental Shift from Positive to Negative Connectivity in Human Amygdala–Prefrontal Circuitry. *Journal of Neuroscience*, *33*(10), 4584–4593. https://doi.org/10.1523/JNEUROSCI.3446-12.2013

Geissberger, N., Tik, M., Sladky, R., Woletz, M., Schuler, A.-L., Willinger, D., & Windischberger, C. (2020). Reproducibility of amygdala activation in facial emotion processing at 7T. *NeuroImage*, *211*, 116585. https://doi.org/10.1016/j.neuroimage.2020.116585

Gelman, A. (2018, March 15). *You need 16 times the sample size to estimate an interaction than to estimate a main effect*. https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models* (1 edition). Cambridge University Press.

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics*, *40*(5), 530–543. https://doi.org/10.3102/1076998615606113

Gelman, A., & Loken, E. (2014). The statistical crisis in science: Data-dependent analysis--a 'garden of forking paths'--explains why many statistically significant comparisons don't hold up. *American Scientist*, *102*(6), 460–466.

Giardino, N. D., Friedman, S. D., & Dager, S. R. (2007). Anxiety, respiration, and cerebral blood flow: Implications for functional brain imaging. *Comprehensive Psychiatry*, *48*(2), 103–112. https://doi.org/10.1016/j.comppsych.2006.11.001

Gilmore, A., Buser, N., & Hanson, J. L. (2020). Variations in Structural MRI Quality Significantly Impact Commonly-Used Measures of Brain Anatomy. *BioRxiv*, 581876. https://doi.org/10.1101/581876

Gitelman, D. R., Penny, W. D., Ashburner, J., & Friston, K. J. (2003). Modeling regional and psychophysiologic interactions in fMRI: The importance of hemodynamic deconvolution. *NeuroImage*, *19*(1), 200–207. https://doi.org/10.1016/S1053-8119(03)00058-2

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron*, *66*(4), 585–595. https://doi.org/10.1016/j.neuron.2010.04.016

Glasser, M. F., Coalson, T. S., Bijsterbosch, J. D., Harrison, S. J., Harms, M. P., Anticevic, A., Van Essen, D. C., & Smith, S. M. (2018). Using temporal ICA to selectively remove global noise while preserving global signal in functional MRI data. *NeuroImage*, *181*, 692–717. https://doi.org/10.1016/j.neuroimage.2018.04.076

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., & Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Glenn, N. D. (2003). Distinguishing Age, Period, and Cohort Effects. In J. T. Mortimer & M. J. Shanahan (Eds.), *Handbook of the Life Course* (pp. 465–476). Springer US. https://doi.org/10.1007/978-0-306-48247-2_21

Glover, G. H., Li, T.-Q., & Ress, D. (2000). Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magnetic Resonance in Medicine*, *44*(1), 162–167. https://doi.org/10.1002/1522-2594(200007)44:1<162::AID-MRM23>3.0.CO;2-E

Golestani, A. M., Chang, C., Kwinta, J. B., Khatamian, Y. B., & Jean Chen, J. (2015). Mapping the end-tidal CO2 response function in the resting-state BOLD fMRI signal: Spatial specificity, test-retest reliability and effect of fMRI sampling rate. *NeuroImage*, *104*, 266–277. https://doi.org/10.1016/j.neuroimage.2014.10.031

Gopnik, A. (2020). Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1803), 20190502. https://doi.org/10.1098/rstb.2019.0502

Gordillo, D., Cruz, J. R. da, Chkonia, E., Lin, W.-H., Favrod, O., Brand, A., Figueiredo, P., Roinishvili, M., & Herzog, M. H. (2022). *The EEG multiverse of schizophrenia* (p. 2020.12.21.20248665). medRxiv. https://doi.org/10.1101/2020.12.21.20248665

Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., Ortega, M., Hoyt-Drazen, C., Gratton, C., Sun, H., Hampton, J. M., Coalson, R. S., Nguyen, A. L., McDermott, K. B., Shimony, J. S., Snyder, A. Z., Schlaggar, B. L., Petersen, S. E., Nelson, S. M., & Dosenbach, N. U. F. (2017). Precision Functional Mapping of Individual Human Brains. *Neuron*, *95*(4), 791-807.e7. https://doi.org/10.1016/j.neuron.2017.07.011

Grady, C. L., Rieck, J. R., Nichol, D., Rodrigue, K. M., & Kennedy, K. M. (2020). Influence of sample size and analytic approach on stability and interpretation of brain-behavior correlations in task-related fMRI data. *Human Brain Mapping*. https://doi.org/10.1002/hbm.25217

Gratton, C., Dworetsky, A., Coalson, R. S., Adeyemo, B., Laumann, T. O., Wig, G. S., Kong, T. S., Gratton, G., Fabiani, M., Barch, D. M., Tranel, D., Miranda-Dominguez, O., Fair, D. A., Dosenbach, N. U. F., Snyder, A. Z., Perlmutter, J. S., Petersen, S. E., & Campbell, M. C. (2020). Removal of high frequency contamination from motion estimates in single-band fMRI saves data without biasing functional connectivity. *BioRxiv*, 837161. https://doi.org/10.1101/837161

Greene, A. S., Gao, S., Noble, S., Scheinost, D., & Constable, R. T. (2020). How Tasks Change Whole-Brain Functional Organization to Reveal Brain-Phenotype Relationships. *Cell Reports*, *32*(8), 108066. https://doi.org/10.1016/j.celrep.2020.108066

Gur, R. C., Sara, R., Hagendoorn, M., Marom, O., Hughett, P., Macy, L., Turner, T., Bajcsy, R., Posner, A., & Gur, R. E. (2002). A method for obtaining 3-dimensional facial expressions and its standardization for use in neurocognitive studies. *Journal of Neuroscience Methods*, *115*(2), 137–143. https://doi.org/10.1016/S0165-0270(02)00006-7

Guyer, A. E., Monk, C. S., McClure-Tone, E. B., Nelson, E. E., Roberson-Nay, R., Adler, A. D., Fromm, S. J., Leibenluft, E., Pine, D. S., & Ernst, M. (2008). A Developmental Examination of Amygdala Response to Facial Expressions. *Journal of Cognitive Neuroscience*, *20*(9), 1565–1582. https://doi.org/10.1162/jocn.2008.20114

Haller, S. P., Chen, G., Kitt, E. R., Smith, A. R., Stoddard, J., Abend, R., Cardenas, S. I., Revzina, O., Coppersmith, D., Leibenluft, E., Brotman, M. A., Pine, D. S., & Pagliaccio,

D. (2022). Reliability of task-evoked neural activation during face-emotion paradigms: Effects of scanner and psychological processes. *Human Brain Mapping*, *n/a*(n/a). https://doi.org/10.1002/hbm.25723

Hare, T. A., Tottenham, N., Galvan, A., Voss, H. U., Glover, G. H., & Casey, B. J. (2008a). Biological Substrates of Emotional Reactivity and Regulation in Adolescence During an Emotional Go-Nogo Task. *Biological Psychiatry*, *63*(10), 927–934. https://doi.org/10.1016/j.biopsych.2008.03.015

Hare, T. A., Tottenham, N., Galvan, A., Voss, H. U., Glover, G. H., & Casey, B. J. (2008b). Biological Substrates of Emotional Reactivity and Regulation in Adolescence During an Emotional Go-Nogo Task. *Biological Psychiatry*, *63*(10), 927–934. https://doi.org/10.1016/j.biopsych.2008.03.015

Hariri, A. R., Tessitore, A., Mattay, V. S., Fera, F., & Weinberger, D. R. (2002). The Amygdala Response to Emotional Stimuli: A Comparison of Faces and Scenes. *NeuroImage*, *17*(1), 317–323. https://doi.org/10.1006/nimg.2002.1179

Hastie, T., Friedman, J., & Tibshirani, R. (2001). Model Assessment and Selection. In T. Hastie, J. Friedman, & R. Tibshirani (Eds.), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (pp. 193–224). Springer. https://doi.org/10.1007/978-0-387-21606-5_7

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random Forests. In *The Elements of Statistical Learning*. https://doi.org/10.1007/978-0-387-84858-7_15

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values ofd'. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51. https://doi.org/10.3758/BF03203619

Heeringa, S. G., Wagner, J., Torres, M., Duan, N., Adams, T., & Berglund, P. (2004). Sample designs and sampling methods for the Collaborative Psychiatric Epidemiology Studies (CPES). *International Journal of Methods in Psychiatric Research*, *13*(4), 221–240. https://doi.org/10.1002/mpr.179

Hein, T. C., Mattson, W. I., Dotterer, H. L., Mitchell, C., Lopez-Duran, N., Thomason, M. E., Peltier, S. J., Welsh, R. C., Hyde, L. W., & Monk, C. S. (2018). Amygdala habituation and uncinate fasciculus connectivity in adolescence: A multi-modal approach. *NeuroImage*, *183*, 617–626. https://doi.org/10.1016/j.neuroimage.2018.08.058

Heller, A. S., Cohen, A. O., Dreyfuss, M. F. W., & Casey, B. J. (2016). Changes in Cortico-Subcortical and Subcortico-Subcortical Connectivity Impact Cognitive Control to Emotional Cues across Development. *Social Cognitive and Affective Neuroscience*, *11*(12), 1910–1918. https://doi.org/10.1093/scan/nsw097

Heller, A. S., Ezie, C. E. C., Otto, A. R., & Timpano, K. R. (2018). Model-based learning and individual differences in depression: The moderating role of stress. *Behaviour Research and Therapy*, *111*, 19–26. https://doi.org/10.1016/j.brat.2018.09.007

Herting, M. M., Gautam, P., Chen, Z., Mezher, A., & Vetter, N. C. (2017). Test-Retest Reliability of Longitudinal Task-Based fMRI: Implications for Developmental Studies. *Developmental Cognitive Neuroscience*. https://doi.org/10.1016/j.dcn.2017.07.001

Hocke, L. M., & Frederick, B. B. (2021). Post-hoc physiological waveform extraction from motion estimation in simultaneous multislice (SMS) functional MRI using separate stack processing. *Magnetic Resonance in Medicine*, *85*(1), 309–315. https://doi.org/10.1002/mrm.28418

Infantolino, Z. P., Luking, K. R., Sauder, C. L., Curtin, J. J., & Hajcak, G. (2018). Robust Is Not Necessarily Reliable: From within-Subjects fMRI Contrasts to between-Subjects Comparisons. *NeuroImage*, *173*, 146–152. https://doi.org/10.1016/j.neuroimage.2018.02.024

Jalbrzikowski, M., Larsen, B., Hallquist, M. N., Foran, W., Calabro, F., & Luna, B. (2017). Development of White Matter Microstructure and Intrinsic Functional Connectivity Between the Amygdala and Ventromedial Prefrontal Cortex: Associations With Anxiety and Depression. *Biological Psychiatry*, *82*(7), 511–521. https://doi.org/10.1016/j.biopsych.2017.01.008

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1006/nimg.2002.1132

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

Jo, H. J., Lee, J.-M., Kim, J.-H., Shin, Y.-W., Kim, I.-Y., Kwon, J. S., & Kim, S. I. (2007). Spatial accuracy of fMRI activation influenced by volume- and surface-based spatial smoothing techniques. *NeuroImage*, *34*(2), 550–564. https://doi.org/10.1016/j.neuroimage.2006.09.047

Johnstone, T., Somerville, L. H., Alexander, A. L., Oakes, T. R., Davidson, R. J., Kalin, N. H., & Whalen, P. J. (2005). Stability of amygdala BOLD response to fearful faces over multiple scan sessions. *NeuroImage*, *25*(4), 1112–1123. https://doi.org/10.1016/j.neuroimage.2004.12.016

Jones, R. M., Somerville, L. H., Li, J., Ruberry, E. J., Libby, V., Glover, G., Voss, H. U., Ballon, D. J., & Casey, B. J. (2011). Behavioral and Neural Properties of Social Reinforcement Learning. *Journal of Neuroscience*, *31*(37), 13039–13045. https://doi.org/10.1523/JNEUROSCI.2972-11.2011

Jones, T. B., Bandettini, P. A., & Birn, R. M. (2008). Integration of motion correction and physiological noise regression in fMRI. *NeuroImage*, *42*(2), 582–590. https://doi.org/10.1016/j.neuroimage.2008.05.019

Jonkman, L. M., Lansbergen, M., & Stauder, J. E. A. (2003). Developmental differences in behavioral and event-related brain responses associated with response preparation and inhibition in a go/nogo task. *Psychophysiology*, *40*(5), 752–761. https://doi.org/10.1111/1469-8986.00075

Joseph, J. E., Zhu, X., Gundran, A., Davies, F., Clark, J. D., Ruble, L., Glaser, P., & Bhatt, R. S. (2015a). Typical and Atypical Neurodevelopment for Face Specialization: An fMRI Study. *Journal of Autism and Developmental Disorders*, *45*(6), 1725–1741. https://doi.org/10.1007/s10803-014-2330-4

Joseph, J. E., Zhu, X., Gundran, A., Davies, F., Clark, J. D., Ruble, L., Glaser, P., & Bhatt, R. S. (2015b). Typical and Atypical Neurodevelopment for Face Specialization: An fMRI Study. *Journal of Autism and Developmental Disorders*, *45*(6), 1725–1741. https://doi.org/10.1007/s10803-014-2330-4

Kaplan, S., Meyer, D., Miranda-Dominguez, O., Perrone, A., Earl, E., Alexopoulos, D., Barch, D. M., Day, T. K. M., Dust, J., Eggebrecht, A. T., Feczko, E., Kardan, O., Kenley, J. K., Rogers, C. E., Wheelock, M. D., Yacoub, E., Rosenberg, M., Elison, J. T., Fair, D. A., & Smyser, C. D. (2022). Filtering respiratory motion artifact from resting state fMRI data in

infant and toddler populations. *NeuroImage*, *247*, 118838. https://doi.org/10.1016/j.neuroimage.2021.118838

Kastrup, A., Krüger, G., Neumann-Haefelin, T., & Moseley, M. E. (2001). Assessment of cerebrovascular reactivity with functional magnetic resonance imaging: Comparison of CO2 and breath holding. *Magnetic Resonance Imaging*, *19*(1), 13–20. https://doi.org/10.1016/S0730-725X(01)00227-2

Kennedy, J. T., Harms, M. P., Korucuoglu, O., Astafiev, S. V., Barch, D. M., Thompson, W. K., Bjork, J. M., & Anokhin, A. P. (2021). *Reliability and Stability Challenges in ABCD Task fMRI Data* (p. 2021.10.08.463750). https://doi.org/10.1101/2021.10.08.463750

Killgore, W. D. S., Oki, M., & Yurgelun-Todd, D. A. (2001). Sex-specific developmental changes in amygdala responses to affective faces. *NeuroReport*, *12*(2), 427.

Killgore, W. D. S., & Yurgelun-Todd, D. A. (2007a). Unconscious Processing of Facial Affect in Children and Adolescents. *Social Neuroscience*, *2*(1), 28–47. https://doi.org/10.1080/17470910701214186

Killgore, W. D. S., & Yurgelun-Todd, D. A. (2007b). Unconscious processing of facial affect in children and adolescents. *Social Neuroscience*, *2*(1), 28–47. https://doi.org/10.1080/17470910701214186

King, K. M., Littlefield, A. K., McCabe, C. J., Mills, K. L., Flournoy, J., & Chassin, L. (2018). Longitudinal modeling in developmental neuroimaging research: Common challenges, and solutions from developmental psychology. *Developmental Cognitive Neuroscience*, *33*, 54–72. https://doi.org/10.1016/j.dcn.2017.11.009

Klapwijk, E., van den Bos, W., Tamnes, C. K., Mills, K. L., & Raschle, N. (2019). *Opportunities for Increased Reproducibility and Replicability of Developmental Cognitive Neuroscience* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/fxjzt

Knudsen, E. I. (2004). Sensitive Periods in the Development of the Brain and Behavior. *Journal of Cognitive Neuroscience*, *16*(8), 1412–1425. https://doi.org/10.1162/0898929042304796

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When Does Model-Based Control Pay Off? *PLOS Computational Biology*, *12*(8), e1005090. https://doi.org/10.1371/journal.pcbi.1005090

Kragel, P. A., Han, X., Kraynak, T. E., Gianaros, P. J., & Wager, T. D. (2021). Functional MRI Can Be Highly Reliable, but It Depends on What You Measure: A Commentary on Elliott et al. (2020). *Psychological Science*, *32*(4), 622–626. https://doi.org/10.1177/0956797621989730

Krebs, R. M., Schott, B. H., Schütze, H., & Düzel, E. (2009). The novelty exploration bonus and its attentional modulation. *Neuropsychologia*, *47*(11), 2272–2281. https://doi.org/10.1016/j.neuropsychologia.2009.01.015

Kriegeskorte, N., Lindquist, M. A., Nichols, T. E., Poldrack, R. A., & Vul, E. (2010). Everything You Never Wanted to Know about Circular Analysis, but Were Afraid to Ask. *Journal of Cerebral Blood Flow & Metabolism*, *30*(9), 1551–1557. https://doi.org/10.1038/jcbfm.2010.86

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. https://doi.org/10.1038/nn.2303

Kujawa, A., Wu, M., Klumpp, H., Pine, D. S., Swain, J. E., Fitzgerald, K. D., Monk, C. S., & Phan, K. L. (2016). Altered Development of Amygdala-Anterior Cingulate Cortex Connectivity in Anxious Youth and Young Adults. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(4), 345–352. https://doi.org/10.1016/j.bpsc.2016.01.006

Kundu, P., Benson, B. E., Baldwin, K. L., Rosen, D., Luh, W.-M., Bandettini, P. A., Pine, D. S., & Ernst, M. (2015). Robust Resting State fMRI Processing for Studies on Typical Brain Development Based on Multi-Echo EPI Acquisition. *Brain Imaging and Behavior*, *9*(1), 56–73. https://doi.org/10.1007/s11682-014-9346-4

Kurz, A. S. (2019, February 10). *Bayesian robust correlations with brms (and why you should love Student's t)*. A. Solomon Kurz. /post/bayesian-robust-correlations-with-brms-and-why-you-should-love-student-s-t/

Lannelongue, L., Grealey, J., & Inouye, M. (2021). Green Algorithms: Quantifying the Carbon Footprint of Computation. *Advanced Science*, *8*(12), 2100707. https://doi.org/10.1002/advs.202100707

Lawrence, R., Loftus, A., Kiar, G., Bridgeford, E. W., Roncal, W. G., Chandrashekhar, V., Mhembere, D., Ryman, S., Zuo, X.-N., Margulies, D. S., Craddock, R. C., Priebe, C. E., Jung, R., Calhoun, V. D., Caffo, B., Burns, R., Milham, M. P., Vogelstein, J. T., & Reproducibility (CoRR), C. for R. and. (2021). *A low-resource reliable pipeline to democratize multi-modal connectome estimation and analysis* (p. 2021.11.01.466686). bioRxiv. https://doi.org/10.1101/2021.11.01.466686

LeDoux, J., & Daw, N. D. (2018). Surviving threats: Neural circuit and computational implications of a new taxonomy of defensive behaviour. *Nature Reviews Neuroscience*, *19*(5), 269–282. https://doi.org/10.1038/nrn.2018.22

Li, X., Ai, L., Giavasis, S., Jin, H., Feczko, E., Xu, T., Clucas, J., Franco, A., Heinsfeld, A. S., Adebimpe, A., Vogelstein, J. T., Yan, C.-G., Esteban, O., Poldrack, R. A., Craddock, C., Fair, D., Satterthwaite, T., Kiar, G., & Milham, M. P. (2021). *Moving Beyond Processing and Analysis-Related Variation in Neuroscience* (p. 2021.12.01.470790). bioRxiv. https://doi.org/10.1101/2021.12.01.470790

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

Liu, Y., Althoff, T., & Heer, J. (2020). Paths Explored, Paths Omitted, Paths Obscured: Decision Points & Selective Reporting in End-to-End Data Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376533

Liu, Y., Kale, A., Althoff, T., & Heer, J. (2021). Boba: Authoring and Visualizing Multiverse Analyses. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1753–1763. https://doi.org/10.1109/TVCG.2020.3028985

Lloyd, A., McKay, R., Sebastian, C. L., & Balsters, J. H. (2021). Are adolescents more optimal decision-makers in novel environments? Examining the benefits of heightened exploration in a patch foraging paradigm. *Developmental Science*, *24*(4), e13075. https://doi.org/10.1111/desc.13075

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

Luna, B., Garver, K. E., Urban, T. A., Lazar, N. A., & Sweeney, J. A. (2004). Maturation of Cognitive Processes From Late Childhood to Adulthood. *Child Development*, *75*(5), 1357–1372. https://doi.org/10.1111/j.1467-8624.2004.00745.x

Luna, B., Tervo-Clemmens, B., & Calabro, F. J. (2021). Considerations When Characterizing Adolescent Neurocognitive Development. *Biological Psychiatry*, *89*(2), 96–98. https://doi.org/10.1016/j.biopsych.2020.04.026

Lynch, C. J., Silver, B. M., Dubin, M. J., Martin, A., Voss, H. U., Jones, R. M., & Power, J. D. (2020). Prevalent and sex-biased breathing patterns modify functional connectivity MRI in young adults. *Nature Communications*, *11*(1), 5290. https://doi.org/10.1038/s41467-020-18974-9

Mackey, S., & Petrides, M. (2014). Architecture and Morphology of the Human Ventromedial Prefrontal Cortex. *The European Journal of Neuroscience*, *40*(5), 2777–2796. https://doi.org/10.1111/ejn.12654

Madhyastha, T., Peverill, M., Koh, N., McCabe, C., Flournoy, J., Mills, K., King, K., Pfeifer, J., & McLaughlin, K. A. (2017). Current Methods and Limitations for Longitudinal fMRI Analysis across Development. *Developmental Cognitive Neuroscience*. https://doi.org/10.1016/j.dcn.2017.11.006

Madhyastha, T., Peverill, M., Koh, N., McCabe, C., Flournoy, J., Mills, K., King, K., Pfeifer, J., & McLaughlin, K. A. (2018). Current methods and limitations for longitudinal fMRI

analysis across development. *Developmental Cognitive Neuroscience*, *33*, 118–128. https://doi.org/10.1016/j.dcn.2017.11.006

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., Moore, L. A., Conan, G., Uriarte, J., … Dosenbach, N. U. F. (2020). Towards Reproducible Brain-Wide Association Studies. *BioRxiv*, 2020.08.21.257758. https://doi.org/10.1101/2020.08.21.257758

Master, S. L., Eckstein, M. K., Gotlieb, N., Dahl, R., Wilbrecht, L., & Collins, A. G. E. (2020). Disentangling the systems contributing to changes in learning during adolescence. *Developmental Cognitive Neuroscience*, *41*, 100732. https://doi.org/10.1016/j.dcn.2019.100732

Masur, P. K. (2021). Understanding the effects of conceptual and analytical choices on 'finding' the privacy paradox: A specification curve analysis of large-scale survey data. *Information, Communication & Society*, *0*(0), 1–19. https://doi.org/10.1080/1369118X.2021.1963460

Masur, P. K. (2022). *Specr* [R]. https://github.com/masurp/specr (Original work published 2019)

McClure, E. B., Monk, C. S., Nelson, E. E., Zarahn, E., Leibenluft, E., Bilder, R. M., Charney, D. S., Ernst, M., & Pine, D. S. (2004). A developmental examination of gender differences in brain engagement during evaluation of threat. *Biological Psychiatry*, *55*(11), 1047–1055. https://doi.org/10.1016/j.biopsych.2004.02.013

McDougle, S. D., Boggess, M. J., Crossley, M. J., Parvin, D., Ivry, R. B., & Taylor, J. A. (2016). Credit assignment in movement-dependent reinforcement learning. *Proceedings of the*

National Academy of Sciences, *113*(24), 6797–6802. https://doi.org/10.1073/pnas.1523669113

McLaren, D. G., Ries, M. L., Xu, G., & Johnson, S. C. (2012). A generalized form of context-dependent psychophysiological interactions (gPPI): A comparison to standard approaches. *NeuroImage*, *61*(4), 1277–1286. https://doi.org/10.1016/j.neuroimage.2012.03.068

Menon, V., Adleman, N. E., White, C. d., Glover, G. h., & Reiss, A. l. (2001). Error-related brain activation during a Go/NoGo response inhibition task. *Human Brain Mapping*, *12*(3), 131–143. https://doi.org/10.1002/1097-0193(200103)12:3<131::AID-HBM1010>3.0.CO;2-C

Meyer, H. C., & Lee, F. S. (2019). Translating Developmental Neuroscience to Understand Risk for Psychiatric Disorders. *American Journal of Psychiatry*, *176*(3), 179–185. https://doi.org/10.1176/appi.ajp.2019.19010091

Milham, M. P., Vogelstein, J., & Xu, T. (2021). Removing the Reliability Bottleneck in Functional Magnetic Resonance Imaging Research to Achieve Clinical Utility. *JAMA Psychiatry*, *78*(6), 587. https://doi.org/10.1001/jamapsychiatry.2020.4272

Miller, K. J., Brody, C. D., & Botvinick, M. M. (2016). Identifying Model-Based and Model-Free Patterns in Behavior on Multi-Step Tasks. *BioRxiv*, 096339. https://doi.org/10.1101/096339

Minsky, M. (1961). Steps toward Artificial Intelligence. *Proceedings of the IRE*, *49*(1), 8–30. https://doi.org/10.1109/JRPROC.1961.287775

Mkrtchian, A., Aylward, J., Dayan, P., Roiser, J. P., & Robinson, O. J. (2017). Modeling Avoidance in Mood and Anxiety Disorders Using Reinforcement Learning. *Biological Psychiatry*, *82*(7), 532–539. https://doi.org/10.1016/j.biopsych.2017.01.017

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680–692. https://doi.org/10.1038/s41562-017-0180-8

Monk, C. S., McClure, E. B., Nelson, E. E., Zarahn, E., Bilder, R. M., Leibenluft, E., Charney, D. S., Ernst, M., & Pine, D. S. (2003). Adolescent Immaturity in Attention-Related Brain Engagement to Emotional Facial Expressions. *NeuroImage*, *20*(1), 420–428. https://doi.org/10.1016/S1053-8119(03)00355-0

Moriceau, S., & Sullivan, R. M. (2006). Maternal presence serves as a switch between learning fear and attraction in infancy. *Nature Neuroscience; New York*, *9*(8), 1004–1006. http://dx.doi.org.ezproxy.cul.columbia.edu/10.1038/nn1733

Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B., & Bandettini, P. A. (2009). The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced? *NeuroImage*, *44*(3), 893–905. https://doi.org/10.1016/j.neuroimage.2008.09.036

Muschelli, J., Gherman, A., Fortin, J.-P., Avants, B., Whitcher, B., Clayden, J. D., Caffo, B. S., & Crainiceanu, C. M. (2019). Neuroconductor: An R platform for medical imaging analysis. *Biostatistics*, *20*(2), 218–239. https://doi.org/10.1093/biostatistics/kxx068

Muschelli, J., Nebel, M. B., Caffo, B. S., Barber, A. D., Pekar, J. J., & Mostofsky, S. H. (2014). Reduction of motion-related artifacts in resting state fMRI using aCompCor. *NeuroImage*, *96*, 22–35. https://doi.org/10.1016/j.neuroimage.2014.03.028

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface*, *14*(134), 20170213. https://doi.org/10.1098/rsif.2017.0213

Nearing, J. T., Douglas, G. M., Hayes, M. G., MacDonald, J., Desai, D. K., Allward, N., Jones, C. M. A., Wright, R. J., Dhanani, A. S., Comeau, A. M., & Langille, M. G. I. (2022). Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*, *13*(1), 342. https://doi.org/10.1038/s41467-022-28034-z

Nee, D. E. (2019). FMRI replicability depends upon sufficient individual-level data. *Communications Biology*, *2*(1), 1–4. https://doi.org/10.1038/s42003-019-0378-6

Nikolaidis, A., Heleniak, C., Fields, A., Bloom, P. A., VanTieghem, M., Vannucci, A., Camacho, N. L., Choy, T., Gibson, L., Harmon, C., Hadis, S. S., Douglas, I. J., Milham, M. P., & Tottenham, N. (2022). Heterogeneity in caregiving-related early adversity: Creating stable dimensions and subtypes. *Development and Psychopathology*, 1–14. https://doi.org/10.1017/S0954579421001668

Nord, C. L., Gray, A., Charpentier, C. J., Robinson, O. J., & Roiser, J. P. (2017). Unreliability of Putative fMRI Biomarkers during Emotional Face Processing. *NeuroImage*, *156*, 119–127. https://doi.org/10.1016/j.neuroimage.2017.05.024

Nussenbaum, K., & Hartley, C. A. (2019). Reinforcement learning across development: What insights can we draw from a decade of research? *Developmental Cognitive Neuroscience*, *40*, 100733. https://doi.org/10.1016/j.dcn.2019.100733

Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., Evans, M. D., & Hartley, C. A. (2020). Moving Developmental Research Online: Comparing In-Lab and Web-Based Studies of

Model-Based Reinforcement Learning. *Collabra: Psychology*, *6*(1). https://doi.org/10.1525/collabra.17213

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron*, *38*(2), 329–337. https://doi.org/10.1016/S0896-6273(03)00169-7

Olsson-Collentine, A., Aert, R. C. M. van, Bakker, M., & Wicherts, J. (2021). *Preprint - Meta-Analyzing the Multiverse: A Peek Under the Hood of Selective Reporting*. PsyArXiv. https://doi.org/10.31234/osf.io/43yae

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, *3*(2), 173–182. https://doi.org/10.1038/s41562-018-0506-1

O'Reilly, J. X., Woolrich, M. W., Behrens, T. E. J., Smith, S. M., & Johansen-Berg, H. (2012). Tools of the Trade: Psychophysiological Interactions and Functional Connectivity. *Social Cognitive and Affective Neuroscience*, *7*(5), 604–609. https://doi.org/10.1093/scan/nss055

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The Curse of Planning: Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive. *Psychological Science*, *24*(5), 751–761. https://doi.org/10.1177/0956797612463080

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences*, *110*(52), 20941–20946. https://doi.org/10.1073/pnas.1312011110

Panda, S., Palaniappan, S., Xiong, J., Bridgeford, E. W., Mehta, R., Shen, C., & Vogelstein, J. T. (2019). hyppo: A Multivariate Hypothesis Testing Python Package. In *ArXiv e-prints*. https://ui.adsabs.harvard.edu/abs/2019arXiv190702088P

Parkes, L., Fulcher, B., Yücel, M., & Fornito, A. (2018). An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. *NeuroImage*, *171*, 415–436. https://doi.org/10.1016/j.neuroimage.2017.12.073

Passarotti, A. M., Sweeney, J. A., & Pavuluri, M. N. (2009). Neural Correlates of Incidental and Directed Facial Emotion Processing in Adolescents and Adults. *Social Cognitive and Affective Neuroscience*, *4*(4), 387–398. https://doi.org/10.1093/scan/nsp029

Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, *68*(9), 1046–1058. https://doi.org/10.1016/j.jclinepi.2015.05.029

Pattwell, S. S., Liston, C., Jing, D., Ninan, I., Yang, R. R., Witztum, J., Murdock, M. H., Dincheva, I., Bath, K. G., Casey, B. J., Deisseroth, K., & Lee, F. S. (2016). Dynamic Changes in Neural Circuitry during Adolescence Are Associated with Persistent Attenuation of Fear Memories. *Nature Communications*, *7*, 11475. https://doi.org/10.1038/ncomms11475

Perlman, S. B., & Pelphrey, K. A. (2011). Developing Connections for Affective Regulation: Age-Related Changes in Emotional Brain Connectivity. *Journal of Experimental Child Psychology*, *108*(3), 607–620. https://doi.org/10.1016/j.jecp.2010.08.006

Pernet, C. R. (2014). Misconceptions in the use of the General Linear Model applied to functional MRI: A tutorial for junior neuro-imagers. *Frontiers in Neuroscience*, *8*, 1. https://doi.org/10.3389/fnins.2014.00001

Pestilli, F. (2022). *Free cloud platform for secure neuroscience data analysis.* brainlife.io. https://github.com/brainlife/brainlife (Original work published 2020)

Pfeifer, J. H., Masten, C. L., Moore, W. E., Oswald, T. M., Mazziotta, J. C., Iacoboni, M., & Dapretto, M. (2011). Entering Adolescence: Resistance to Peer Influence, Risky Behavior, and Neural Changes in Emotion Reactivity. *Neuron*, *69*(5). https://doi.org/10.1016/j.neuron.2011.02.019

Phelps, E. A., Delgado, M. R., Nearing, K. I., & LeDoux, J. E. (2004). Extinction Learning in Humans: Role of the Amygdala and vmPFC. *Neuron*, *43*(6), 897–905. https://doi.org/10.1016/j.neuron.2004.08.042

Pine, D. S., Grun, J., Zarahn, E., Fyer, A., Koda, V., Li, W., Szeszko, P. R., Ardekani, B., & Bilder, R. M. (2001). Cortical Brain Regions Engaged by Masked Emotional Faces in Adolescents and Adults: An fMRI Study. *Emotion*, *1*(2), 137–147. https://doi.org/10.1037/1528-3542.1.2.137

Plichta, M. M., Grimm, O., Morgen, K., Mier, D., Sauer, C., Haddad, L., Tost, H., Esslinger, C., Kirsch, P., Schwarz, A. J., & Meyer-Lindenberg, A. (2014). Amygdala habituation: A reliable fMRI phenotype. *NeuroImage*, *103*, 383–390. https://doi.org/10.1016/j.neuroimage.2014.09.059

Poskanzer, C., Fang, M., Aglinskas, A., & Anzellotti, S. (2021). Controlling for Spurious Nonlinear Dependence in Connectivity Analyses. *Neuroinformatics*. https://doi.org/10.1007/s12021-021-09540-9

Potter, T. C. S., Bryce, N. V., & Hartley, C. A. (2017). Cognitive components underpinning the development of model-based learning. *Developmental Cognitive Neuroscience*, *25*, 272–280. https://doi.org/10.1016/j.dcn.2016.10.005

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., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2013). Steps toward optimizing motion artifact removal in functional connectivity MRI; a reply to Carp. *NeuroImage*, *76*, 10.1016/j.neuroimage.2012.03.017. https://doi.org/10.1016/j.neuroimage.2012.03.017

Power, J. D., Lynch, C. J., Adeyemo, B., & Petersen, S. E. (2020). A Critical, Event-Related Appraisal of Denoising in Resting-State fMRI Studies. *Cerebral Cortex*. https://doi.org/10.1093/cercor/bhaa139

Power, J. D., Lynch, C. J., Silver, B. M., Dubin, M. J., Martin, A., & Jones, R. M. (2019). Distinctions among real and apparent respiratory motions in human fMRI data. *NeuroImage*, *201*, 116041. https://doi.org/10.1016/j.neuroimage.2019.116041

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, *84*, 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048

Power, J. D., Plitt, M., Gotts, S. J., Kundu, P., Voon, V., Bandettini, P. A., & Martin, A. (2018). Ridding fMRI data of motion-related influences: Removal of signals with distinct spatial

and physical bases in multiecho data. *Proceedings of the National Academy of Sciences*, *115*(9), E2105–E2114. https://doi.org/10.1073/pnas.1720985115

Power, J. D., Plitt, M., Laumann, T. O., & Martin, A. (2017). Sources and implications of whole-brain fMRI signals in humans. *NeuroImage*, *146*, 609–625. https://doi.org/10.1016/j.neuroimage.2016.09.038

Pozzi, E., Vijayakumar, N., Rakesh, D., & Whittle, S. (2020). Neural correlates of emotion regulation in adolescents and emerging adults: A meta-analytic study. *Biological Psychiatry*, *0*(0). https://doi.org/10.1016/j.biopsych.2020.08.006

Prokopiou, P. C., Pattinson, K. T. S., Wise, R. G., & Mitsis, G. D. (2019). Modeling of dynamic cerebrovascular reactivity to spontaneous and externally induced CO2 fluctuations in the human brain using BOLD-fMRI. *NeuroImage*, *186*, 533–548. https://doi.org/10.1016/j.neuroimage.2018.10.084

Pruim, R. H. R., Mennes, M., Buitelaar, J. K., & Beckmann, C. F. (2015). Evaluation of ICA-AROMA and alternative strategies for motion artifact removal in resting state fMRI. *NeuroImage*, *112*, 278–287. https://doi.org/10.1016/j.neuroimage.2015.02.063

Qin, S., Young, C. B., Duan, X., Chen, T., Supekar, K., & Menon, V. (2014). Amygdala Subregional Structure and Intrinsic Functional Connectivity Predicts Individual Differences in Anxiety During Early Childhood. *Biological Psychiatry*, *75*(11), 892–900. https://doi.org/10.1016/j.biopsych.2013.10.006

Raab, H. A., & Hartley, C. A. (2018). Chapter 13—The Development of Goal-Directed Decision-Making. In R. Morris, A. Bornstein, & A. Shenhav (Eds.), *Goal-Directed Decision Making* (pp. 279–308). Academic Press. https://doi.org/10.1016/B978-0-12-812098-9.00013-9

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, *92*(437), 179–191. https://doi.org/10.1080/01621459.1997.10473615

Raj, D., Anderson, A. W., & Gore, J. C. (2001). Respiratory effects in human functional magnetic resonance imaging due to bulk susceptibility changes. *Physics in Medicine and Biology*, *46*(12), 3331–3340. https://doi.org/10.1088/0031-9155/46/12/318

Raj, D., Paley, D. P., Anderson, A. W., Kennan, R. P., & Gore, J. C. (2000). A model for susceptibility artefacts from respiration in functional echo-planar magnetic resonance imaging. *Physics in Medicine and Biology*, *45*(12), 3809–3820. https://doi.org/10.1088/0031-9155/45/12/321

Raval, V., Nguyen, K. P., Pinho, M., Dewey, R. B., Trivedi, M., & Montillo, A. A. (2021). *Pitfalls and recommended strategies and metrics for suppressing motion artifacts in functional MRI* (p. 2021.09.18.460908). bioRxiv. https://doi.org/10.1101/2021.09.18.460908

Rescorla, R. A., & Wagner, A. R. (1972). *3 A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement*.

Richeson, J. A., Todd, A. R., Trawalter, S., & Baird, A. A. (2008). Eye-Gaze Direction Modulates Race-Related Amygdala Activity. *Group Processes & Intergroup Relations*, *11*(2), 233–246. https://doi.org/10.1177/1368430207088040

Roesch, M. R., Esber, G. R., Li, J., Daw, N. D., & Schoenbaum, G. (2012). Surprise! Neural Correlates of Pearce-Hall and Rescorla-Wagner Coexist within the Brain. *The European Journal of Neuroscience*, *35*(7), 1190–1200. https://doi.org/10.1111/j.1460-9568.2011.07986.x

Rohrer, J. M., Egloff, B., & Schmukle, S. C. (2017). Probing Birth-Order Effects on Narrow Traits Using Specification-Curve Analysis. *Psychological Science*, *28*(12), 1821–1832. https://doi.org/10.1177/0956797617723726

Rorden, C., Bonilha, L., Fridriksson, J., Bender, B., & Karnath, H.-O. (2012). Age-specific CT and MRI templates for spatial normalization. *NeuroImage*, *61*(4), 957–965. https://doi.org/10.1016/j.neuroimage.2012.03.020

Rovee, C. K., & Rovee, D. T. (1969). Conjugate reinforcement of infant exploratory behavior. *Journal of Experimental Child Psychology*, *8*(1), 33–39. https://doi.org/10.1016/0022-0965(69)90025-3

Salas, J. A., Bayrak, R. G., Huo, Y., & Chang, C. (2021). Reconstruction of respiratory variation signals from fMRI data. *NeuroImage*, *225*, 117459. https://doi.org/10.1016/j.neuroimage.2020.117459

Saltelli, A., & Annoni, P. (2010). How to avoid a perfunctory sensitivity analysis. *Environmental Modelling & Software*, *25*(12), 1508–1517. https://doi.org/10.1016/j.envsoft.2010.04.012

Satterthwaite, T. D., Wolf, D. H., Loughead, J., Ruparel, K., Elliott, M. A., Hakonarson, H., Gur, R. C., & Gur, R. E. (2012). Impact of in-scanner head motion on multiple measures of functional connectivity: Relevance for studies of neurodevelopment in youth. *NeuroImage*, *60*(1), 623–632. https://doi.org/10.1016/j.neuroimage.2011.12.063

Satterthwaite, T. D., Wolf, D. H., Ruparel, K., Erus, G., Elliott, M. A., Eickhoff, S. B., Gennatas, E. D., Jackson, C., Prabhakaran, K., Smith, A., Hakonarson, H., Verma, R., Davatzikos, C., Gur, R. E., & Gur, R. C. (2013). Heterogeneous Impact of Motion on Fundamental Patterns of Developmental Changes in Functional Connectivity during Youth. *NeuroImage*, *83*, 45–57. https://doi.org/10.1016/j.neuroimage.2013.06.045

Sauder, C. L., Hajcak, G., Angstadt, M., & Phan, K. L. (2013). Test-retest reliability of amygdala response to emotional faces. *Psychophysiology*, *50*(11), 1147–1156. https://doi.org/10.1111/psyp.12129

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

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Sergerie, K., Chochol, C., & Armony, J. L. (2008). The role of the amygdala in emotional processing: A quantitative meta-analysis of functional neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, *32*(4), 811–830. https://doi.org/10.1016/j.neubiorev.2007.12.002

Shahar, N., Moran, R., Hauser, T. U., Kievit, R. A., McNamee, D., Moutoussis, M., Consortium, N., & Dolan, R. J. (2019). Credit assignment to state-independent task representations and its relationship with model-based decision making. *Proceedings of the National Academy of Sciences*, *116*(32), 15871–15876. https://doi.org/10.1073/pnas.1821647116

Sharp, M. E., Foerde, K., Daw, N. D., & Shohamy, D. (2016). Dopamine selectively remediates 'model-based' reward learning: A computational approach. *Brain*, *139*(2), 355–364. https://doi.org/10.1093/brain/awv347

Shou, H., Eloyan, A., Lee, S., Zipunnikov, V., Crainiceanu, A. N., Nebel, M. B., Caffo, B., Lindquist, M. A., & Crainiceanu, C. M. (2013). Quantifying the reliability of image replication studies: The image intraclass correlation coefficient (I2C2). *Cognitive,*

Affective, & Behavioral Neuroscience, *13*(4), 714–724. https://doi.org/10.3758/s13415-013-0196-0

Siegel, J. S., Power, J. D., Dubis, J. W., Vogel, A. C., Church, J. A., Schlaggar, B. L., & Petersen, S. E. (2014). Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. *Human Brain Mapping*, *35*(5), 1981–1996. https://doi.org/10.1002/hbm.22307

Silva, C. F. da, Yao, Y.-W., & Hare, T. A. (2018). Can model-free reinforcement learning operate over information stored in working-memory? *BioRxiv*, 107698. https://doi.org/10.1101/107698

Silvers, J. A., Insel, C., Powers, A., Franz, P., Helion, C., Martin, R. E., Weber, J., Mischel, W., Casey, B. J., & Ochsner, K. N. (2017a). VlPFC–vmPFC–Amygdala Interactions Underlie Age-Related Differences in Cognitive Regulation of Emotion. *Cerebral Cortex (New York, NY)*, *27*(7), 3502–3514. https://doi.org/10.1093/cercor/bhw073

Silvers, J. A., Insel, C., Powers, A., Franz, P., Helion, C., Martin, R., Weber, J., Mischel, W., Casey, B. J., & Ochsner, K. N. (2017b). The Transition from Childhood to Adolescence Is Marked by a General Decrease in Amygdala Reactivity and an Affect-Specific Ventral-to-Dorsal Shift in Medial Prefrontal Recruitment. *Developmental Cognitive Neuroscience*, *25*, 128–137. https://doi.org/10.1016/j.dcn.2016.06.005

Silvers, J. A., Shu, J., Hubbard, A. D., Weber, J., & Ochsner, K. N. (2015). Concurrent and Lasting Effects of Emotion Regulation on Amygdala Response in Adolescence and Young Adulthood. *Developmental Science*, *18*(5), 771–784. https://doi.org/10.1111/desc.12260

Simon, D. A., & Daw, N. D. (2011). Environmental statistics and the trade-off between model-based and TD learning in humans. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 24* (pp. 127–135). Curran Associates, Inc. http://papers.nips.cc/paper/4243-environmental-statistics-and-the-trade-off-between-model-based-and-td-learning-in-humans.pdf

Simonsohn, U., Simmons, J., & Nelson, L. (2015). Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. *Marketing Papers*. https://doi.org/10.2139/ssrn.2694998

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Smid, C. R., Kool, W., Hauser, T. U., & Steinbeis, N. (2020). *Model-based decision-making and its metacontrol in childhood* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/ervsb

Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D., & Dolan, R. J. (2013). Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans. *Neuron*, *80*(4), 914–919. https://doi.org/10.1016/j.neuron.2013.08.009

Somerville, L. H., Hare, T., & Casey, B. J. (2010). Frontostriatal Maturation Predicts Cognitive Control Failure to Appetitive Cues in Adolescents. *Journal of Cognitive Neuroscience*, *23*(9), 2123–2134. https://doi.org/10.1162/jocn.2010.21572

Spence, M. J., & DeCasper, A. J. (1987). Prenatal experience with low-frequency maternal-voice sounds influence neonatal perception of maternal voice samples. *Infant Behavior and Development*, *10*(2), 133–142. https://doi.org/10.1016/0163-6383(87)90028-2

Spohrs, J., Bosch, J. E., Dommes, L., Beschoner, P., Stingl, J. C., Geiser, F., Schneider, K., Breitfeld, J., & Viviani, R. (2018). Repeated fMRI in measuring the activation of the amygdala without habituation when viewing faces displaying negative emotions. *PLOS ONE*, *13*(6), e0198244. https://doi.org/10.1371/journal.pone.0198244

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Suess, W. M., Alexander, A. B., Smith, D. D., Sweeney, H. W., & Marion, R. J. (1980). The Effects of Psychological Stress on Respiration: A Preliminary Study of Anxiety and Hyperventilation. *Psychophysiology*, *17*(6), 535–540. https://doi.org/10.1111/j.1469-8986.1980.tb02293.x

Sullivan, R. M., & Perry, R. E. (2015). Mechanisms and Functional Implications of Social Buffering in Infants: Lessons from Animal Models. *Social Neuroscience*, *10*(5), 500–511. https://doi.org/10.1080/17470919.2015.1087425

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction* (1st Edition edition). A Bradford Book.

Swartz, J. R., Carrasco, M., Wiggins, J. L., Thomason, M. E., & Monk, C. S. (2014). Age-Related Changes in the Structure and Function of Prefrontal Cortex–Amygdala Circuitry

in Children and Adolescents: A Multi-Modal Imaging Approach. *NeuroImage*, *86*, 212–220. https://doi.org/10.1016/j.neuroimage.2013.08.018

Telzer, E. H., Flannery, J., Humphreys, K. L., Goff, B., Gabard-Durman, L., Gee, D. G., & Tottenham, N. (2015). "The Cooties Effect": Amygdala Reactivity to Opposite-versus Same-Sex Faces Declines from Childhood to Adolescence. *Journal of Cognitive Neuroscience*, *27*(9), 1685–1696. https://doi.org/10.1162/jocn_a_00813

Telzer, E. H., Humphreys, K. L., Shapiro, M., & Tottenham, N. (2012). Amygdala Sensitivity to Race Is Not Present in Childhood but Emerges over Adolescence. *Journal of Cognitive Neuroscience*, *25*(2), 234–244. https://doi.org/10.1162/jocn_a_00311

Telzer, E. H., McCormick, E. M., Peters, S., Cosme, D., Pfeifer, J. H., & van Duijvenvoorde, A. C. K. (2018). Methodological Considerations for Developmental Longitudinal fMRI Research. *Developmental Cognitive Neuroscience*. https://doi.org/10.1016/j.dcn.2018.02.004

Teruel, J. R., Kuperman, J. M., Dale, A. M., & White, N. S. (2018). High temporal resolution motion estimation using a self-navigated simultaneous multi-slice echo planar imaging acquisition. *Journal of Magnetic Resonance Imaging: JMRI*. https://doi.org/10.1002/jmri.25953

Thomas, K. M., Drevets, W. C., Dahl, R. E., Ryan, N. D., Birmaher, B., Eccard, C. H., Axelson, D., Whalen, P. J., & Casey, B. J. (2001). Amygdala response to fearful faces in anxious and depressed children. *Archives of General Psychiatry*, *58*(11), 1057–1063.

Thompson, G. J., Merritt, M. D., Pan, W.-J., Magnuson, M. E., Grooms, J. K., Jaeger, D., & Keilholz, S. D. (2013). Neural correlates of time-varying functional connectivity in the rat. *NeuroImage*, *83*, 826–836. https://doi.org/10.1016/j.neuroimage.2013.07.036

Tijssen, R. H. N., Jenkinson, M., Brooks, J. C. W., Jezzard, P., & Miller, K. L. (2014). Optimizing RetroICor and RetroKCor corrections for multi-shot 3D FMRI acquisitions. *NeuroImage*, *84*, 394–405. https://doi.org/10.1016/j.neuroimage.2013.08.062

Tobe, R. H., MacKay-Brandt, A., Lim, R., Kramer, M., Breland, M. M., Trautman, K. D., Hu, C., Sangoi, R., Tu, L., Alexander, L., Gabbay, V., Castellanos, F. X., Leventhal, B. L., Craddock, R. C., Colcombe, S. J., Franco, A. R., & Milham, M. P. (2021). *A Longitudinal Resource for Studying Connectome Development and its Psychiatric Associations During Childhood* (p. 2021.03.09.21253168). https://doi.org/10.1101/2021.03.09.21253168

Todd, R. M., Evans, J. W., Morris, D., Lewis, M. D., & Taylor, M. J. (2011). The Changing Face of Emotion: Age-Related Patterns of Amygdala Activation to Salient Faces. *Social Cognitive and Affective Neuroscience*, *6*(1), 12–23. https://doi.org/10.1093/scan/nsq007

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, B., & Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, *168*(3), 242–249. https://doi.org/10.1016/j.psychres.2008.05.006

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., Hazlewood, V., Lathrop, S., Lifka, D., Peterson, G. D., Roskies, R., Scott, J. R., & Wilkins-Diehr, N. (2014). XSEDE: Accelerating Scientific Discovery. *Computing in Science Engineering*, *16*(5), 62–74. https://doi.org/10.1109/MCSE.2014.80

Tustison, N. J., Avants, B. B., Cook, P. A., Song, G., Das, S., Strien, N. van, Stone, J. R., & Gee, J. C. (2013). The ANTs cortical thickness processing pipeline. *Medical Imaging 2013:*

Biomedical Applications in Molecular, Structural, and Functional Imaging, *8672*, 86720K. https://doi.org/10.1117/12.2007128

US Census Bureau. (2021). *U.S. Census Bureau QuickFacts: Los Angeles County, California*. https://www.census.gov/quickfacts/losangelescountycalifornia

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E. J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., Della Penna, S., Feinberg, D., Glasser, M. F., Harel, N., Heath, A. C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., … WU-Minn HCP Consortium. (2012). The Human Connectome Project: A data acquisition perspective. *NeuroImage*, *62*(4), 2222–2231. https://doi.org/10.1016/j.neuroimage.2012.02.018

Vanderwal, T., Eilbott, J., & Castellanos, F. X. (2019). Movies in the magnet: Naturalistic paradigms in developmental functional neuroimaging. *Developmental Cognitive Neuroscience*, *36*, 100600. https://doi.org/10.1016/j.dcn.2018.10.004

VanTieghem, M., Korom, M., Flannery, J., Choy, T., Caldera, C., Humphreys, K. L., Gabard-Durnam, L., Goff, B., Gee, D. G., Telzer, E. H., Shapiro, M., Louie, J. Y., Fareri, D. S., Bolger, N., & Tottenham, N. (2021). Longitudinal changes in amygdala, hippocampus and cortisol development following early caregiving adversity. *Developmental Cognitive Neuroscience*, *48*, 100916. https://doi.org/10.1016/j.dcn.2021.100916

Varendi, H., Porter, R. H., & Winberg, J. (1996). Attractiveness of amniotic fluid odor: Evidence of prenatal olfactory learning? *Acta Paediatrica*, *85*(10), 1223–1227. https://doi.org/10.1111/j.1651-2227.1996.tb18233.x

Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R., & Andreou, P. (2013). The Case for Using the Repeatability Coefficient When Calculating Test–Retest Reliability. *PLOS ONE*, *8*(9), e73990. https://doi.org/10.1371/journal.pone.0073990

Vazire, S., & Holcombe, A. O. (2021). Where are the Self-Correcting Mechanisms in Science? *Review of General Psychology*, 10892680211033912. https://doi.org/10.1177/10892680211033912

Vijayakumar, N., Pfeifer, J. H., Flournoy, J. C., Hernandez, L. M., & Dapretto, M. (2019). Affective reactivity during adolescence: Associations with age, puberty and testosterone. *Cortex*, *117*, 336–350. https://doi.org/10.1016/j.cortex.2019.04.024

Vincent, T., Ciuciu, P., & Thirion, B. (2008). Sensitivity analysis of parcellation in the joint detection-estimation of brain activity in fMRI. *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 568–571. https://doi.org/10.1109/ISBI.2008.4541059

Vink, M., Derks, J. M., Hoogendam, J. M., Hillegers, M., & Kahn, R. S. (2014). Functional Differences in Emotion Processing during Adolescence and Early Adulthood. *NeuroImage*, *91*, 70–76. https://doi.org/10.1016/j.neuroimage.2014.01.035

Voegtline, K. M., Costigan, K. A., Pater, H. A., & DiPietro, J. A. (2013). Near-term fetal response to maternal spoken voice. *Infant Behavior and Development*, *36*(4), 526–533. https://doi.org/10.1016/j.infbeh.2013.05.002

Wager, T. D., & Nichols, T. E. (2003). Optimization of experimental design in fMRI: A general framework using a genetic algorithm. *NeuroImage*, *18*(2), 293–309.

Wallis, L. A., Healy, M., Undy, M. B., & Maconochie, I. (2005). Age related reference ranges for respiration rate and heart rate from 4 to 16 years. *Archives of Disease in Childhood*, *90*(11), 1117–1121. https://doi.org/10.1136/adc.2004.068718

Wang, Y., & Luo, Y. (2005). Standardization and Assessment of College Students' Facial Expression of Emotion. *Chinese Journal of Clinical Psychology*, *13*(4), 396–398.

Weermeijer, J., Lafit, G., Kiekens, G., Wampers, M., Eisele, G., Kasanova, Z., Vaessen, T., Kuppens, P., & Myin-Germeys, I. (2022). Applying multiverse analysis to experience sampling data: Investigating whether preprocessing choices affect robustness of conclusions. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01777-1

Werchan, D. M., Collins, A. G. E., Frank, M. J., & Amso, D. (2015). 8-Month-Old Infants Spontaneously Learn and Generalize Hierarchical Rules. *Psychological Science*, *26*(6), 805–815. https://doi.org/10.1177/0956797615571442

Werchan, D. M., Collins, A. G. E., Frank, M. J., & Amso, D. (2016). Role of Prefrontal Cortex in Learning and Generalizing Hierarchical Rules in 8-Month-Old Infants. *Journal of Neuroscience*, *36*(40), 10314–10322. https://doi.org/10.1523/JNEUROSCI.1351-16.2016

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, *7*. https://www.frontiersin.org/article/10.3389/fpsyg.2016.01832

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019).

Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wong, C. W., Olafsson, V., Tal, O., & Liu, T. T. (2013). The amplitude of the resting-state fMRI global signal is related to EEG vigilance measures. *NeuroImage*, *83*, 983–990. https://doi.org/10.1016/j.neuroimage.2013.07.057

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *NeuroImage*, *14*(6), 1370–1386. https://doi.org/10.1006/nimg.2001.0931

Wu, M., Kujawa, A., Lu, L. H., Fitzgerald, D. A., Klumpp, H., Fitzgerald, K. D., Monk, C. S., & Phan, K. L. (2016). Age-related changes in amygdala–frontal connectivity during emotional face processing from childhood into young adulthood. *Human Brain Mapping*, *37*(5), 1684–1695. https://doi.org/10.1002/hbm.23129

Xifra-Porxas, A., Kassinopoulos, M., & Mitsis, G. D. (2021). Physiological and motion signatures in static and time-varying functional connectivity and their subject identifiability. *ELife*, *10*, e62324. https://doi.org/10.7554/eLife.62324

Xu, J., Hao, L., Chen, M., He, Y., Jiang, M., Tian, T., Wang, H., Wang, Y., Wang, D., Han, Z. R., Tan, S., Men, W., Gao, J., He, Y., Tao, S., Dong, Q., & Qin, S. (2021). Developmental Sex Differences in Negative Emotion Decision-Making Dynamics: Computational Evidence and Amygdala-Prefrontal Pathways. *Cerebral Cortex*, bhab359. https://doi.org/10.1093/cercor/bhab359

Xu, T., Cho, J. W., Kiar, G., Bridgeford, E. W., Vogelstein, J. T., & Milham, M. P. (2022). *A Guide for Quantifying and Optimizing Measurement Reliability for the Study of*

Individual Differences (p. 2022.01.27.478100). bioRxiv. https://doi.org/10.1101/2022.01.27.478100

Yan, C.-G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R. C., Di Martino, A., Li, Q., Zuo, X.-N., Castellanos, F. X., & Milham, M. P. (2013). A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. *NeuroImage*, *76*, 183–201. https://doi.org/10.1016/j.neuroimage.2013.03.004

Yousefi, B., Shin, J., Schumacher, E. H., & Keilholz, S. D. (2018). Quasi-periodic patterns of intrinsic brain activity in individuals and their relationship to global signal. *NeuroImage*, *167*, 297–308. https://doi.org/10.1016/j.neuroimage.2017.11.043

Yurgelun-Todd, D. A., & Killgore, W. D. S. (2006a). Fear-related activity in the prefrontal cortex increases with age during adolescence: A preliminary fMRI study. *Neuroscience Letters*, *406*(3), 194–199. https://doi.org/10.1016/j.neulet.2006.07.046

Yurgelun-Todd, D. A., & Killgore, W. D. S. (2006b). Fear-Related Activity in the Prefrontal Cortex Increases with Age during Adolescence: A Preliminary fMRI Study. *Neuroscience Letters*, *406*(3), 194–199. https://doi.org/10.1016/j.neulet.2006.07.046

Zhang, Y., Padmanabhan, A., Gross, J. J., & Menon, V. (2019). Development of human emotion circuits investigated using a Big-Data analytic approach: Stability, reliability, and robustness. *Journal of Neuroscience*, 0220–19. https://doi.org/10.1523/JNEUROSCI.0220-19.2019

Zhu, D. C., Tarumi, T., Khan, M. A., & Zhang, R. (2015). Vascular Coupling in Resting-State FMRI: Evidence from Multiple Modalities. *Journal of Cerebral Blood Flow & Metabolism*, *35*(12), 1910–1920. https://doi.org/10.1038/jcbfm.2015.166

Zuo, X.-N., Xu, T., & Milham, M. P. (2019). Harnessing reliability for neuroscience research. *Nature Human Behaviour*, *3*(8), 768–771. https://doi.org/10.1038/s41562-019-0655-x

Appendix A: Chapter 1 Supplement

Previous Studies of Age-Related Change in Amygdala—mPFC Function

Authors	Year	N	Design	Ages	Task	Faces	MRI Design	Contrast	Reactivity	Amyg—mPFC Connectivity	Connectivity Method
Baird et al.	1999	12	Cross-sectional	12 to 17	Emotion labeling with fear faces	Ekman & Friesen, 1976	Block	fear > nonsense grayscale figures	No age-related changes	NA	NA
Killgore et al.	2001	19	Cross-sectional	9 to 17	Emotion labeling with fear faces	Ekman & Friesen, 1976	Block	fear > baseline	Age-related decrease in left amyg, but not right amyg, only in females	NA	NA
Thomas et al.	2001	18	Cross-sectional	Youth (mean = 11, sd = 2.4), and male adults (mean = 24, sd = 6.6)	Passive viewing with fear & neutral faces	Ekman & Friesen, 1976	Block	fear > neutral	Adults show higher fear > neutral reactivity in left amyg than children	NA	NA
Pine et al.	2001	20	Cross-sectional	Youth (age 12-16) and adults (age 25-38)	Masking paradigm with happy and fear faces	Ekman & Friesen, 1976	Block	masked fear > fixation, masked happy > fixation	No group differences in amygdala reactivity for any contrast	NA	NA
Monk et al.	2003	34	Cross-sectional	Youth (9-17) and adults (25-36)	1) passive viewing with fear, angry, happy, neutral 2) emotional rating of faces 3) subjective rating of nose width	NimStim (2009); Ekman & Friesen, 1976; Gur, 2001	Event-related	A) rating of fear for fear faces > nose width for fear faces, B) nose width for fear faces > nose width for neutral faces, C) passive viewing for fear faces > passive viewing for neutral faces	No age-related amygdala differences for contrasts A/B. For C, adolescents show higher reactivity than adults in right amyg, but no age-related change within adolescent group	NA	NA

Authors	Year	N	Design	Ages	Task	Faces	MRI Design	Contrast	Reactivity	Amyg—mPFC Connectivity	Connectivity Method
McClure et al.	2004	34	Cross-sectional	Youth (9-17) and adults (25-36)	Threat rating for fear, angry, happy, neutral faces	NimStim (2009); Ekman & Friesen, 1976; Gur, 2001	Event-related	Angry faces > all other faces, each emotion separately > baseline	Only in females, adults showed greater right amygdala reactivity to angry > neutral and angry > fear faces	NA	NA
Yurgelun-Todd & Killgore	2006	16	Cross-sectional	8 to 15	Passive viewing of fearful and happy faces	Ekman & Friesen, 1976	Block	fear > baseline, happy > baseline	No age-related changes in amygdala reactivity for either contrast	NA	NA
Killgore & Yurgelun-Todd	2007	22	Cross-sectional	Adolescents (age 9-17) and adults (mean= 23.7, sd=2.1)	Masking paradigm with sad and happy faces	Erwin et al., 1992	Block	masked sad > baseline, masked happy > baseline, masked happy > masked sad	Adolescents show greater right amygdala activation for masked sad > baseline than adults. No differences for other contrasts	NA	NA
Guyer et al.	2008	61	Cross-sectional	Adolescents (9-17) and adults (21-40)	Passive viewing with fear, angry, happy, neutral faces	NimStim (2009); Ekman & Friesen, 1976; Gur, 2000	Event-related	fear > neutral, fear > fixation, neutral > fixation	Adolescents show greater amygdala activation than adults for fear > neutral and fear > fixation. No difference for the neutral > fixation contrast. No age-related change within adolescent group.	No differences in seed-based amyg—mPFC functional connectivity between adolescents and adults	Seed-based correlation
Hare et al.	2008	60	Cross-sectional	Children (7-12), Adolescents (13-18), Adults (19-32)	Go/no-go with fear, happy, calm faces. All combinations of emotions used as targets	NimStim (2009)	Event-related	fear > baseline, fear > calm	Adolescents show greatest amygdala activation to fear > baseline compared to children and adults. No differences in amygdala activation to fear > baseline between children and adults	NA	NA
Passarotti et al.	2009	20	Cross-sectional	Adolescents (mean age = 14), adults (mean age 30)	Age judgement, affect judgement with angry and happy faces	Gur, 2002	Event-related	angry + happy > fixation	Adolescents show greater right amyg activation than adults in the incidental condition > fixation under a liberal contiguity threshold, but not a strict one. No group differences in amygdala during directed condition	NA	NA

Authors	Year	N	Design	Ages	Task	Faces	MRI Design	Contrast	Reactivity	Amyg—mPFC Connectivity	Connectivity Method
Somerville et al.	2010	62	Cross-sectional	6 to 29	Go/no-go with fear, happy, calm faces. All combinations of emotions used as targets	NimStim (2009)	Event-related	happy > baseline, calm > baseline	No age-related changes in amygdala from whole-brain analysis	NA	NA
Pfeifer et al	2011	38 (76 scans)	Longitudinal (2 waves)	10 to 13	Passive viewing with fear, angry, happy, sad, neutral faces	NimStim (2009)	Event-related	all faces individually > fixation, all faces together > fixation, all emotional faces individually > neutral faces	No age-related change in right or left amygdala activity in all faces averaged together > fixation. Increase in activity in the right amygdala in response to sad > neutral faces	NA	NA
Forbes et al.	2011	76	Cross-sectional	11 to 13	Matching with fear, angry, neutral faces, and shapes	NimStim (2009)	Block	fear > shapes, angry > shapes, neutral > shapes	No differences for fear > shapes. Pre/early-pubertal adolescents show greater amyg reactivity to neutral > shapes than mid/late pubertal adolescents	NA	NA
Todd et al.	2011	45	Cross-sectional	31 children (age 3.5-8.5), 14 young adults (age 18-33)	Passive viewing of angry and happy mother/stranger faces, phase-scrambled images	Angry & happy face images of mothers of participants	Block	all faces > scrambled faces, angry > scrambled faces, happy > scrambled faces	Age-related increase in bilateral amyg activation to faces > scrambled faces, and angry faces > scrambled faces	NA	NA
Perlman & Pelphrey	2011	20	Cross-sectional	5 to 11	Go/no-go with fear faces interspersed. Also, block structure with winning, losing, and recovery of points	NimStim (2009)	Event-related	Effective connectivity calculated during block 3 of the task	NA	Effective connectivity during block 3 between left amygdala and inferior frontal gyrus/ACC increased with age	Granger causality

Authors	Year	N	Design	Ages	Task	Faces	MRI Design	Contrast	Reactivity	Amyg—mPFC Connectivity	Connectivity Method
Telzer et al.	2012	32	Cross-sectional	4 to 16.5	Matching with angry, happy, neutral faces, and shapes	NimStim (2009)	Block	African-American faces > baseline, European-American faces > baseline	Right amygdala response to African-American faces > baseline increases with age, no age-related change in right amygdala response to European-American faces	NA	NA
Gee et al.	2013	45	Cross-sectional	4 to 22	Go/no-go with fear, happy, and neutral faces. Withhold press for emotional faces	NimStim (2009)	Event-related	fear > baseline	Age-related decreases in right amygdala	Positive early in development between right amyg—mPFC, turning negative in adolescence	gPPI with AFNI deconvolution
Swartz et al.	2014	39	Cross-sectional	9 to 19	Gender identification with fear, happy, sad, neutral faces	NimStim (2009)	Event-related	all emotions individually > baseline	Age-related decreases in left amygdala to fear > baseline. Age-related decreases in left amygdala to each other emotion individually > baseline	No age-related change in right or left amygdala with mPFC for all faces > baseline contrast	gPPI with SPM8, deconvolution not mentioned
Dreyfuss et al.	2014	80	Cross-sectional	6 to 27	Go/no-go with fear, happy, calm faces. All combinations of emotions used as targets	NimStim (2009)	Event-related	fear > calm	No age-related changes in amygdala from whole-brain analysis	NA	NA
Telzer et al.	2015	52	Cross-sectional	4 to 18	Matching with angry, happy, neutral, faces, and shapes	NimStim (2009)	Block	same sex > shapes, opposite sex > shapes	Bilateral amygdala responses to opposite-sex faces > shapes decrease with age, bilateral amygdala response to same-sex faces > shapes increase with age	NA	NA
Joseph et al.	2015	42	Cross-sectional	5 to 18	Passive viewing of faces (73% happy, the rest neutral),	Custom unfamiliar high-school yearbook faces	Block	face > texture	Age-related increase in right and left amyg reactivity to faces > textures contrast	NA	NA

Authors	Year	N	Design	Ages	Task	Faces	MRI Design	Contrast	Reactivity	Amyg—mPFC Connectivity	Connectivity Method
					objects, and textures						
Wu et al.	2016	61	Cross-sectional	7 to 25	Matching with fear, angry, happy, neutral faces, and shapes.	Hariri et al., 2002	Block	fear > shapes, angry > shapes, happy > shapes	No age-related change in amygdala found for any contrast, or all together	For each emotion > shapes, positive early in development between left amyg-ACC, turning to negative around age 15. Same with right amyg	PPI done in AFNI
Kujawa et al.	2016	61	Cross-sectional	7 to 25	Matching with fear, angry, happy, neutral faces, and shapes.	Hariri et al., 2002	Block	fear > shapes, angry > shapes, happy > shapes	No age-related change effects in either hemisphere for either emotion > shapes	Positive early in development between both hemispheres/ACC, turning negative in around age 15	PPI done in SPM with deconvolution
Heller et al.	2016	155	Cross-sectional	5 to 32	Go/no-go with fear, happy, calm faces. All combinations of emotions used as targets	NimStim (2009)	Event-related	happy > baseline, calm > baseline	NA	No age-related changes reported in whole-brain analysis	Beta series correlation analysis done with amygdala seed to whole brain
Vijayakumar et al.	2019	82	Longitudinal (3 waves)	9 to 18	Passive viewing with fear, angry, happy, sad, neutral faces	NimStim (2009)	Event-related	all emotions together > baseline, all emotions individually > baseline	None for fear > baseline. Age-related increase for sad > baseline, and sad > neutral	NA	NA
Zhang et al.	2019	759	Cross-Sectional	8 to 23	Emotion labeling with fear, angry, happy, sad faces	Gur, 2002	Rapid Event-related	all emotions individually > baseline	Depending on ROI definition and model, either no age-related change or age-related increases for fear > baseline, happy > baseline, sad > baseline	Depending on ROI definition and model, either no age-related change in amyg-PFC connectivity, or age-related increases for fear, happy, angry	gPPI & BSC
Xu et al.	2021	321	Cross-Sectional	243 (7-12), 78 young adults (19-25)	Emotion matching with fear & angry faces	Wang & Luo (2005)	Block	negative (fear/angry) faces > shapes	Age*sex interaction -- age-related decreases in females in BLA and CMA, increases (though not always significant) in males	age*sex interaction -- age-related increases in males, decreases in females	gPPI with deconvolution using SPM

Appendix A Table 1: Previous studies of age-related change in amygdala—mPFC function

We summarize cohort, task design, contrast, connectivity method (when available), and amygdala—mPFC result information from prior work with fMRI analyses of age-related differences in amygdala—mPFC responses to faces (Baird et al., 1999; Dreyfuss et al., 2014; Ekman & Friesen, 1976; Erwin et al., 1992; Forbes et al., 2011; Gee et al., 2013; Gur et al., 2002; Guyer et al., 2008; Hare et al., 2008b; Hariri et al., 2002; Heller et al., 2016; Joseph et al., 2015b; Killgore et al., 2001; Killgore & Yurgelun-Todd, 2007b; Kujawa et al., 2016; McClure et al., 2004; Monk et al., 2003; Passarotti et al., 2009; Perlman & Pelphrey, 2011; Pfeifer et al., 2011; Pine et al., 2001; Somerville et al., 2010; Swartz et al., 2014; Telzer et al., 2012, 2015; Thomas et al., 2001; Todd et al., 2011; Tottenham et al., 2009; Vijayakumar et al., 2019; Wang & Luo, 2005; Wu et al., 2016; J. Xu et al., 2021; Yurgelun-Todd & Killgore, 2006b; Zhang et al., 2019).

Supplemental Methods

Participant demographics

Parents reported their child's gender, race, and whether their child was *"Hispanic or Latino"* via questionnaire items (except for 10 participants ages 18-22 who self-reported). Parents indicated separately whether their child's race fell in differeint categories (*African-American/Black, American Indian/Alaska Native, Asian-American, European-American/Caucasian, Native Hawaiian or Other Pacific Islander;* see Appendix A Table 2). Some parents indicated multiple race categories (therefore total proportions of racial groups sum to greater than 1). Parents also had the option to indicate any other racial categories their children belonged to through an open-response item (see Appendix A Table 3). Parents also reported annual household income within discrete bins (see Appendix A Figure 1). Income data was missing for 13 families.

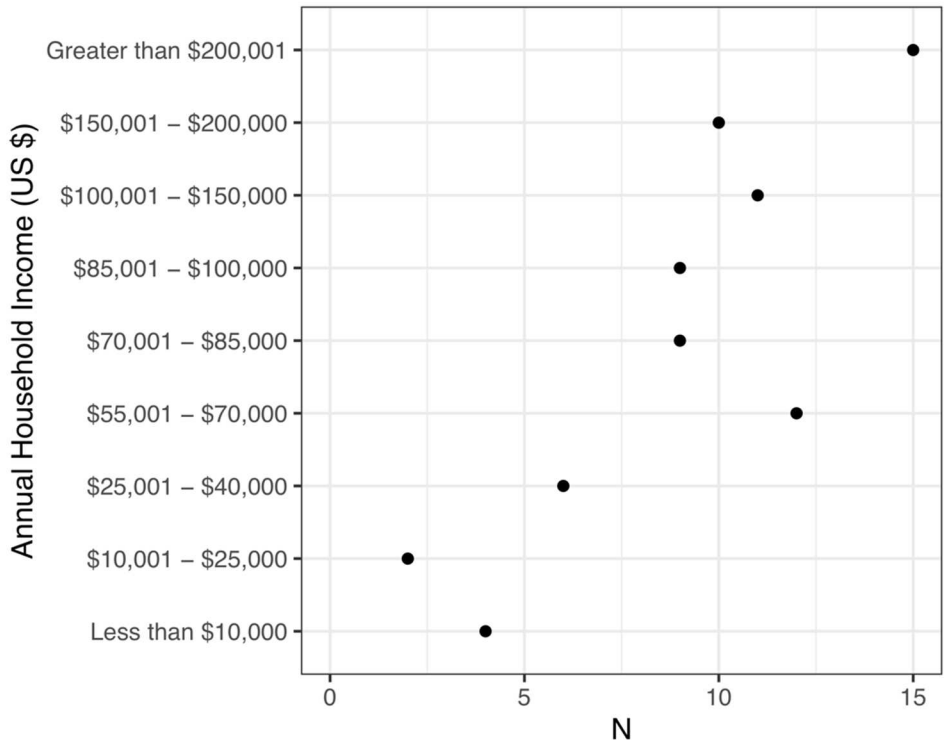
		N	*Proportion*
Gender	Female	55	0.56

	Male	43	0.44
Race	European-American/Caucasian	56	0.57
	African-American/Black	28	0.29
	Asian-American	24	0.24
	American Indian/Alaska Native	3	0.03
	Native Hawaiian or Other Pacific Islander	1	0.01
Hispanic or Latino ethnicity	Hispanic or Latino	12	0.12
	Not Hispanic or Latino	84	0.86
	Missing Data	2	0.02

Appendix A Table 2: Race, gender, and ethnicity distributions for study participants

Response	*N*
Hispanic	3
Arab	2
Filipino	1
Iranian American	1
Mexican	1
Spanish	1

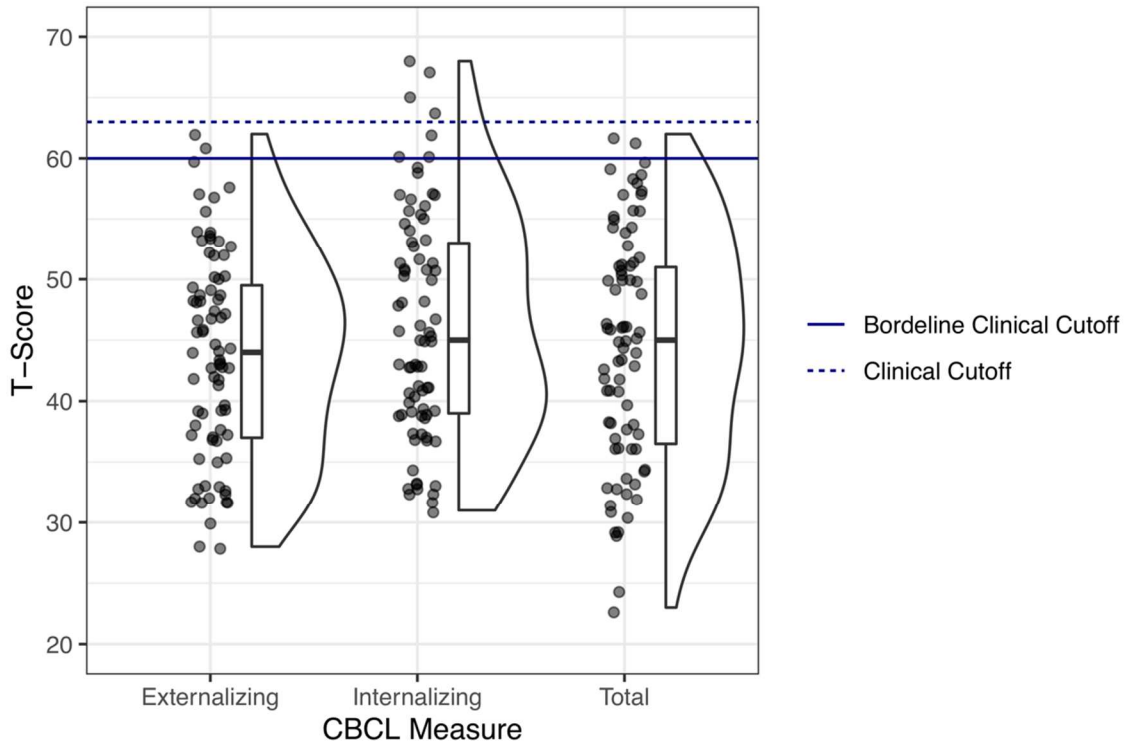
Appendix A Table 3: Parent free responses for their children's racial identities beyond available categories



Appendix A Figure 1: Distribution of annual household income for participating families.

CBCL Scores:

Parents reported on child emotional and behavioral symptoms via the Child Behavior Checklist (CBCL; Achenbach, 1991) for age 1.5-5 and 4-18 years. Age and gender-normed scores for participants' first study timepoint indicated that 4 participants met clinical threshold for internalizing problems, while none met clinical threshold for externalizing or total problems.



Appendix A Figure 2: Parent-reported CBCL scores

Boxplots show medians and IQRs for CBCL T-scores for each scale, and points are individual participants. No participants met clinical cutoff for externalizing or total problems at timepoint 1, while 4 participants met the clinical cutoff for internalizing problems.

Analyses of task behavior

To characterize age-related changes in task performance, we modeled behavior in several ways. First, we calculated the d' performance metric for each scan using a correction for extreme proportions (Hautus, 1995) where button presses for neutral faces were considered 'hits', withheld presses for neutral faces 'misses', withheld presses for fear faces 'correct rejections', and presses for fear faces 'false alarms'. We constructed multilevel models to look at linear only,

linear + quadratic, and linear + quadratic + cubic age-related changes in d' with the following R syntax:

```
# Linear:
brm(dprime ~ age + (age | participant))

# Linear + Quadratic:
brm(dprime ~ poly(age, 2, raw = TRUE) + (age | participant))

# Linear + Quadratic + Cubic:
brm(dprime ~ poly(age, 3, raw = TRUE) + (age | participant))

# Inverse age (1/age)
brm(dprime ~ age_inverse + (age_inverse | participant))
```

We fit all models using the brms package (Bürkner, 2017), and all models included varying linear slopes for age and intercepts across participants. In addition, to characterize more specific aspects of task behavior, we fit separate single-trial multilevel logistic regression models with overall accuracy (probability of a correct response on any trial), hit rate (on neutral face trials), and false alarm rate (on fear face trials) as the respective outcomes and included nested varying

effects for task sessions within participants. For the accuracy model, ‘hits’ and ‘correct rejections’ were coded as 1 and ‘misses’ and ‘false alarms’ coded as 0.

```
# Accuracy:
brm(accuracy ~ age + (age |participant/session), family = bernoulli(link = 'logit'))

# False Alarms:
brm(false_alarms  ~ age + (age |participant/session), family = bernoulli(link = 'logit') )

# Hits:
brm(hits  ~ age + (age |participant/session), family = bernoulli(link = 'logit'))
```

To examine age-related changes in reaction times (for hits only), we used similar single-trial multilevel regression models with linear only, linear + quadratic, and linear + quadratic + cubic terms and nested varying effects for task sessions within participants.

```
# Linear:
brm(reaction_time ~ age + (age |participant / session))

# Linear + Quadratic:
brm(reaction_time ~ poly(age, 2, raw = TRUE) + (age |participant / session))

# Linear + Quadratic + Cubic:
```

brm(reaction_time ~ poly(age, 3, raw = TRUE) + (age |participant / session))

Inverse Age (age_inverse = 1/age)

brm(reaction_time ~ age_inverse + (age_inverse |participant / session))

Reactivity Analyses

C-PAC preprocessing pipeline

In addition to the preregistered FSL preprocessing pipeline, we preprocessed BOLD images using C-PAC software (v1.4.1; Craddock et al., 2013). For these pipelines, we mostly used software defaults. In these pipelines, ANTS (Tustison et al., 2013) was used to skull-strip MPRAGE images, and slice-time correction was applied (slices were acquired in interleaved order). BOLD spatial realignment and motion parameters were calculated with MCFLIRT as with the preregistered pipeline. Registration matrices were then calculated for functional images to be registered to high-resolution structural T1 images using FSL's FLIRT with boundary-based registration. Registration matrices for T1 images to standard MNI space were calculated using ANTS, and functional images were warped to MNI space before running GLMs.

Amygdala reactivity: multiverse details (Table 2, Aim 1)

As detailed in in the main manuscript, we constructed forking pipelines for analyses of age-related change in amygdala reactivity. Below we provide details into each decision point.

- GLM Software: Within-participant first-level GLMs were conducted either using FEAT or AFNI 3dDeconvolve. Prewhitening as specified in software defaults was used in all GLMs to adjust for temporal autocorrelation.
- Hemodynamic Response Function: Within the GLM, regressors for fear and neutral faces were convolved with either a canonical double-gamma or single-gamma hemodynamic response function.
- Nuisance Regressors: Nuisance regressors were added to the first-level GLMs in all analysis pipelines. Pipelines included either 6 head motion regressors, 24 motion regressors (as preregistered; 3 for rotation, 3 for translation, their temporal derivatives, and the square roots of all the above; see Friston et al., 1996), or 18 motion regressors plus additional regressors for mean white matter and cerebrospinal fluid signal (C-PAC preprocessing only, 6 head motion regressors, their squares, and their backwards derivatives). In addition, to remove low-frequency artifacts, we either applied a high-pass filter (cutoff = .01Hz) to BOLD data before the GLM, or included a quadratic drift term in the model (AFNI GLMs only). In all pipelines, TRs with >.9mm framewise displacement (FD) were down-weighted to 0 in the GLM, effectively removing their influence on the model.
- First-Level GLM Estimates: From each first-level GLM, we estimated contrasts for fear faces > baseline, neutral faces > baseline, and fear > neutral faces for each voxel using an event-related design. Although the fear > baseline contrast was of primary interest to the current study in following up work by Gee et al. (2013), we repeated analyses for the other contrasts as well. For each contrast, we either submitted beta estimates for each contrast (i.e. FSL COPEs), or t-statistics for group-level models. While the beta estimates

represent the raw magnitude of estimated contrast effects for each scan, the t-statistics represent standardized effect sizes; i.e. these magnitudes scaled by the estimation uncertainty.

- Amygdala ROIs: We conducted reactivity analyses with the amygdala defined in both native space and in standard MNI space. For native space analyses, participant-specific native space masks were defined using Freesurfer (v6.0; Fischl, 2012) for the bilateral amygdala, as well as left and right amygdala separately, and manually inspected for quality assurance by an experimenter. For analyses in standard space, we used an amygdala mask from the Harvard-Oxford Atlas (probability threshold = .5). In addition, to check whether effects were driven by signal dropout, we performed a median split of all voxels in the right and left amygdala, respectively, based on the grand mean BOLD intensity across all scans to create 'high signal' and 'low signal' amygdala sub-regions. All amygdala masks are available on OSF (https://osf.io/hvdmx/). For all amygdala ROI definitions, we calculated the mean reactivity across included voxels for each scan for bilateral, left, and right regions respectively for group-level analyses.
- Exclusion of previously analyzed scans: 45 of the scans in this dataset were previously analyzed by Gee et al. (2013). Because these scans were used in a whole-brain mass univariate analysis as a discovery sample to identify regions changing with age in both reactivity and connectivity with the amygdala for the fear > baseline contrast, analyses in the current study including these scans will could be partially dependent (i.e. circular analysis) on the previous selection process (Kriegeskorte et al., 2009, 2010). Therefore, we included pipelines in the multiverse that excluded the 42 scans (3 scans originally included were excluded in the current study for high motion) previously analyzed. In

addition, to disentangle whether any differences in results between such pipelines are due to nonindependence versus reduced sample size (i.e. reduced estimation precision), we also conducted permutation testing in which we iteratively removed 42 scans *not* originally analyzed in the Gee et al. study before group-level modeling.

- Outliers: As preregistered, scans where amygdala reactivity estimates were > 3 standard deviations from the mean were excluded from analysis. In addition, to account for the possibility of remaining outliers, additional models were fit using a Student's t-distribution (rather than a Gaussian) for the likelihood function. A gamma(4,1) prior was used for the parameter for degrees of freedom (ν). Such models, because they model the outcome variable with a heavy-tailed t-distributions (i.e. t-distributions with few degrees of freedom), have been demonstrated to be robust to outliers (Gelman & Hill, 2006; Kurz, 2019).
- Group-level models: ROI estimates from each scan were submitted to multilevel linear regression models to estimate group-level effects. Age was grand mean-centered and modeled as a continuous variable, and all models included a covariate for mean framewise displacement (Power et al., 2012). In multiverse analyses, we included models with all combinations of additional covariates for task run (coded as a binary variable indicating first run versus second/third run) and scanner (coded as a binary variable for scanner 1 versus scanner 2). In addition, we included models with an additional quadratic term for age to explore potential non-linear age-related changes. We included models both with and without participant-specific random slope terms, but all models included participant-specific intercepts.

- Within-participant change vs. between-participant differences: To ask whether any observed age-related changes in amygdala reactivity were due to true developmental growth (i.e. longitudinal change within the same participant across timepoints) versus between-participant differences, we conducted a separate smaller multiverse analysis. In this analysis, we included separate model parameters for 'within-participant' (i.e. mean-centered within participants) and 'between-participant' (i.e. grand mean-centered) age effects, with random effects for within-participant age. These analyses included preprocessing pipelines using both FSL and C-PAC, GLMs run in both FSL and AFNI, and both native and MNI space amygdala t-stat estimates.

Longitudinal Amygdala Model Syntax

As described above, we modeled longitudinal age-related changes in amygdala reactivity using several different specifications with the brms package. R syntax for the 9 specifications for longitudinal models is shown below. Each model (plus equivalent models with normal likelihood functions instead) was applied to all 156 preprocessing pipelines for a total of 2808 models.

```
# 1: Linear, no exclusions
modLinear = brm(reactivity ~ age_centered + motion + (age_centered|participant), data = ., cores = 2, chains = 4, family = 'student', prior = prior(gamma(4, 1), class = nu))

# 2: Linear, exclude previously studied participants
```

```
modLinearExclude = brm(reactivity ~ age_centered + motion + (age_centered|participant), data
= dplyr::filter(., is.na(prev_studied)), cores = 2, chains = 4, family = 'student', prior =
prior(gamma(4, 1), class = nu))

# 3: Quadratic, exclude previously studied participants
modQuadraticExclude = brm(reactivity ~ poly(age_centered,2, raw = TRUE) + motion +
(age_centered|participant), data = dplyr::filter(., is.na(prev_studied)), cores = 2, chains = 4,
family = 'student', prior = prior(gamma(4, 1), class = nu))

# 4: Quadratic, no exclusions
modQuadratic = brm(reactivity ~ poly(age_centered,2, raw = TRUE) + motion +
(age_centered|participant), data = ., cores = 2, chains = 4, family = 'student', prior =
prior(gamma(4, 1), class = nu))

# 5: Linear + scanner covariate, no exclusions
modLinearScanner = brm(reactivity ~ age_centered + motion + scanner +
(age_centered|participant), data = ., cores = 2, chains = 4, family = 'student', prior =
prior(gamma(4, 1), class = nu))

# 6: Linear + block covariate, no exclusions
modLinearBlock = brm(reactivity ~ age_centered + motion + blockBin +
(age_centered|participant), data = ., cores = 2, chains = 4, family = 'student', prior =
prior(gamma(4, 1), class = nu))
```

```
# 7: Quadratic + block + scanner covariates, no exclusions
modQuadraticBlockScanner = brm(reactivity ~ poly(age_centered,2, raw = TRUE) + motion +
blockBin + scanner + (age_centered|participant), data = ., cores = 2, chains = 4, family =
'student', prior = prior(gamma(4, 1), class = nu))

# 8: Linear + block + scanner covariates, no exclusions
modLinearBlockScanner = brm(reactivity ~ age_centered + motion + blockBin + scanner +
(age_centered|participant), data = ., cores = 2, chains = 4, family = 'student', prior =
prior(gamma(4, 1), class = nu)))

# 9: Linear without participant-varying slopes, no exclusions
modLinearNoRandomSlopes = brm(reactivity ~ age_centered + motion + (1|participant), data =
., cores = 2, chains = 4, family = 'student', prior = prior(gamma(4, 1), class = nu)))
```

Parametrization of within-participant change versus between-participant differences

In efforts to better discriminate truly longitudinal within-participant changes from between-participant differences, we created alternate model parametrizations for a subset of specifications (bilateral amygdala only, t-statistic amygdala reactivity estimates) for both the fear > baseline and neutral > baseline contrast. For each model, we included two separate terms for age. The first term represented between-participant differences and was grand mean-centered

(such that this term was equal to 0 at age 11.9 years; the mean age across all participants and scans); the second, representing within-participant change, was centered within participants (such that this term was equal to 0 at the mean age of each participant across visits). We allowed only the second within-participant change term to vary across participants, as follows:

```
# Within vs. between model (reactivity)
brm(reactivity ~ age_grand_mean_centered + age_centered_within_participant +
        motion + (age_centered_within_participant|participant))
```

After fitting these model formulations to the subset of preprocessing specifications, we examined the posterior distributions of resulting grand mean-centered (between-participant) and within-participant centered (within-participant) parameters in a smaller specification curve (Figure 2D). In addition, we also examined approximate leave-one-out cross-validated R^2 using the loo_R2() function from the brms package (Bürkner, 2017). This metric uses the posterior likelihood to provide an adjusted R^2 metric that estimates predictive performance.

Within-person similarity of reactivity voxel-wise statistical maps across multiverse forks

We asked whether, for each given scan, whether different pipelines yielded similar voxel-wise patterns of estimates in a within-scan analysis. To explore how changes in preprocessing impacted scan-level statistical results for the fear > baseline, neutral > baseline, and fear > neutral contrasts, we computed image similarities between t-statistic maps for each scan across all pipelines. Similarities were calculated using product-moment correlations of 3d images (using AFNI 3ddot; https://afni.nimh.nih.gov/pub/dist/doc/program_help/3ddot.html) for the same scan after transformation to MNI space at 2mm (isotropic) resolution, and across all voxels in the

brain. We computed similarities across all pairwise comparisons of pipelines preprocessed using C-PAC, as well the similarity of all C-PAC pipelines to the preregistered FSL pipeline. In addition to whole-brain similarity, we also computed similarity statistics in the same way for all voxels within the bilateral amygdala mask from the Harvard Oxford atlas.

Between-scan associations between amygdala reactivity estimates across specifications

We also asked whether different pipelines yielded similar relationships between amygdala betas across scans in a between-scan analysis. This analysis examined whether between-scan relationships among amygdala reactivity estimates were preserved across preprocessing specifications for each contrast. To accomplish this, we computed rank-order correlations between amygdala-reactivity estimates (t-tstats, bilateral amygdala only) between preprocessing specifications for each contrast. Because this analysis was focused on examining whether scan-level differences were preserved across specifications, correlations were conducted across all scans from all participants without taking into account the nesting of repeated observations within participants across sessions.

Dependence of amygdala reactivity findings on previous work

42 scans from the first timepoint were previously analyzed as a 'discovery set' by Gee et al (2013). Thus, we were concerned that analyses of amygdala reactivity including all data (including these scans) might be biased to find stronger effects of negative age-related change due to the pre-selection of the amygdala for showing an effect within part of the current sample (see Kriegeskorte et al., 2009). Indeed, age-related change in amygdala reactivity for the fear > baseline contrast was stronger and more negative in specifications using all data, compared to

when excluding these 42 scans (see Appendix A Figure 25). However, such stronger effects without exclusion may also have been due to the larger sample size. To address this, we conducted permutation tests across several pipelines to assess differences in age-related change estimates after exclusion of these 42 scans versus 42 other randomly-drawn scans. To conserve computational resources, we fit these models using maximum likelihood with the lme4 R package (Bates et al., 2019), rather than fully Bayesian inference. Approximate 95% confidence intervals were constructed from these models by computing the interval ±2 standard errors from the maximum likelihood estimate.

Within-scan changes in amygdala reactivity across trials

Changes in amygdala reactivity across trials: multiverse details (Table 2, Aim 2)

- Quantification of change across trials: We conducted separate multiverse analyses using several different methods for measuring change in reactivity across trials.
- *Slopes:* For each scan we performed a rank-order correlation between trial number and amygdala betas corresponding to each trial number (separately for fear and neutral trials). These correlation coefficients representing slopes for linear changes in reactivity across trials were then submitted to group-level models.
- *Trial halves:* We split trials into the first half (trials 1-12) and second half (trials 13-24) for fear and neutral faces respectively, and modeled age-related change in amygdala reactivity separately for each half of trials at the group level. These models included half x age interaction terms to specifically estimate whether age-related change in amygdala reactivity differed in the first versus second half of trials.

- *Single-trial models:* We constructed single-trial models with age x trial number interaction terms and crossed random effects for age (random intercept and age effects for each participant) and trial number (random intercepts and linear slopes for trial for each scan).
- Global signal subtraction: It is possible that changes in amygdala reactivity across trials may reflect temporal trends in the whole-brain 'global signal', rather than differing amygdala reactivity specifically. To correct for this, we included pipelines with a global signal correction using post-hoc distribution centering of reactivity for each trial. Post-hoc distribution centering consisted of subtracting the mean beta estimate across all voxels from each voxel such that the distribution of beta estimates for each respective trial was centered at 0 (mean-centering).
- Amygdala ROI: Only Harvard-Oxford anatomically-defined amygdala ROIs in MNI space were used in these analyses. Analyses of bilateral, left, and right amygdala ROIs were each included.
- Group-level models: Age was grand mean-centered and modeled as a continuous variable, and all models included a covariate for mean framewise displacement. In multiverse analyses, we included models with all combinations of additional covariates for task run (coded as a binary variable indicating first run versus second/third run) and scanner (coded as a binary variable for scanner 1 versus scanner 2). In addition, we included models with an additional quadratic term for age. All models were formulated to be robust to outliers as described above.

Example brms model formulas are shown below for each of the different types of models for within-scan changes in amygdala reactivity.

```
# Slope model (method 1)
slope_model_ = brm(slope ~ age_centered + motion + (age_centered | participant),
      data = ., cores = 2, chains =4, family = 'student',
      prior = prior(gamma(4, 1), class = nu))

# Trial half model (method 2)
half_model = brm(reactivity ~ age_centered*half + motion +
(age_centered|participant) + (1 | scan),
      data = ., cores = 2, chains =4, family = 'student',
      prior = prior(gamma(4, 1), class = nu))

# Single trial model (linear trial term, method 3)
single_trial_linear = brm(reactivity ~ age_centered*trial +
      motion + (age_centered | participant ) + (trial | scan),
      data = ., cores = 2, chains =4, family = 'student',
      prior = prior(gamma(4, 1), class = nu))

# Single trial model (discrete trial term, method 3)
single_trial_discrete = brm(reactivity ~ age_centered*trial_discrete +
      motion + (age_centered | participant ) + (trial | scan),
```

```
data = ., cores = 2, chains =4, family = 'student',
prior = prior(gamma(4, 1), class = nu))
```

Amygdala—mPFC functional connectivity analyses

Amygdala—mPFC gPPI analyses: multiverse details (Table 2, Aim 3)

- Preprocessing: We ran all gPPI analyses using preregistered preprocessing pipelines in FSL. No gPPI analyses were run on data preprocessed using C-PAC.
- Amygdala gPPI seed: All gPPI analyses used the anatomically-defined Harvard-Oxford bilateral amygdala mask as a seed region. We extracted mean timecourses from the amygdala from the preprocessed BOLD data to use as the seed regressor.
- Deconvolution step: In all pipelines interaction terms between the amygdala seed timeseries and both fear and neutral face regressors were constructed following generalized form (gPPI; McLaren et al., 2012). Before multiplication of the seed and stimulus timeseries, however, some pipelines included an additional step such that the seed timeseries was deconvolved to recreate the seed 'neuronal' timeseries using AFNI 3dTfitter (https://afni.nimh.nih.gov/pub/dist/doc/program_help/3dTfitter.html). To most closely match pipelines run by Gee et al. (2013), we did not up-sample the seed timeseries before using 3dTfitter, and applied no regularization to the regression solver. This deconvolution step (often included in AFNI and SPM PPI analyses, but not FSL) has been suggested to allow the PPI regressors to better approximate task-modulated connectivity at the level of the neuronal response, rather than after being filtered by the hemodynamic response function (Di & Biswal, 2017; Gitelman et al., 2003). Following

deconvolution, the resulting timeseries was multiplied with the fear and neutral face regressors, then re-convolved with the HRF for entry into the GLM. For pipelines not including a deconvolution step, seed timeseries were multiplied by stimulus regressors that had already been convolved with the HRF.

- First-Level gPPI GLM: First-level GLMs for gPPI analyses were constructed in FSL almost identically to those used in preregistered amygdala reactivity analyses, with the addition of the amygdala seed timeseries regressor and gPPI terms for both fear and neutral faces. gPPI GLMs included 24 head motion parameters and had TRs with framewise displacement >.9mm down-weighted to 0. As with amygdala reactivity analyses, we extracted both t-statistics and contrast beta estimates (COPEs) for the fear > baseline, neutral > baseline, and fear > neutral contrasts, although the fear > baseline contrast was of primary interest for the current study.
- mPFC ROI definition: We preregistered constructing an mPFC ROI containing 120 voxels centered at the peak coordinates reported by Gee at al. (2013) for age-related change in fear > baseline gPPI (Tailarach 2,32,8; or MNI 3,35,8). However, after discovery that this ROI heavily overlapped the corpus collosum, we instead constructed three spherical ROIs with 5mm radii, the first centered at the above peak coordinates, the second shifted slightly anterior, and the third shifted slightly ventral relative to the second (see Figure 4). Lastly, we also used a large mask encompassing the 'whole vmPFC', taken from Mackey & Petrides (2014). All masks used are available on OSF (https://osf.io/hvdmx/). For each scan, we calculated mean gPPI beta estimates and t-statistics for each of these four ROIs for submission to group-level models.

- Group-level models, outliers, and previously analyzed scans: Multiverse gPPI analyses included identical decision points to amygdala reactivity analyses with respect to group-level modeling, dealing with outliers, and use of previously analyzed scans.

Amygdala—mPFC BSC analyses: multiverse details (Table 2, Aim 3)

- Preprocessing: We conducted all BSC analyses using preregistered preprocessing pipelines in FSL.
- GLMs: We used beta estimates from LSS GLMs fit separately to each trial (as described on p. 12 of the main text) for BSC analyses.
- Amygdala ROI definitions: Only Harvard-Oxford amygdala ROIs in MNI space were used in these analyses. Analyses of bilateral, left, and right amygdala ROIs were each included.
- mPFC ROI definitions: For BSC analyses, we used the same mPFC ROIs as with gPPI analyses (see Table 2, Aim 3).
- Global signal subtraction: We included BSC pipelines both with and without the global signal subtraction step previously described (post-hoc distribution centering).
- Beta-series correlations: For each respective amygdala—mPFC ROI pair (12 pairs in total for 3 amygdala x 4 mPFC), we extracted mean beta estimates for each ROI for each trial, then calculated product-moment correlations between the timeseries across trials (neutral and fear separately) for both regions (Di et al., 2020). These correlation coefficients were then submitted to group-level models.
- Group-level models: Multiverse BSC analyses included identical decision points to gPPI analyses with respect to group-level modeling, with the exceptions that all BSC models

were formulated to be robust to outliers (using Student's t likelihood functions), and we did not exclude previously analyzed scans from any BSC analysis pipelines.

Longitudinal models for functional connectivity

Longitudinal models for gPPI used the same 9 specifications as previously described (with R syntax) above for amygdala reactivity. BSC models were similar, except that all BSC models used t-distributed likelihood functions and no BSC models excluded previously studied participants because BSC analyses had not previously been conducted with these data. Thus, there were a total of 7 longitudinal model specifications for BSC analyses.

Rank-order associations between gPPI & BSC amygdala—mPFC FC

In addition to examining age-related changes in amygdala—mPFC FC using both gPPI and BSC methods, we asked whether between-scan differences in scan-level FC estimates were similar across methods. To accomplish this, we computed rank-order correlations between amygdala—mPFC FC estimates using both gPPI (with deconvolution and without) and BSC (with global signal subtraction and without) methods for all four mPFC ROIs. Because this analysis aimed to investigate the scan-level similarity of FC estimates across method, we computed correlations across all scans for all participants.

Effects of up-sampling/lasso regularization parameters during deconvolution on gPPI regressors

The AFNI documentation for the 3dTfitter program comes with the warning, "Deconvolution is a tricky business, so be careful out there! ... Experiment with different parameters to make sure the results in your type of problems make sense"

(https://afni.nimh.nih.gov/pub/dist/doc/program_help/3dTfitter.html). While our initial choices of deconvolution parameters were guided by previous work done with the same dataset, we explored the impact of different parameters on the resulting gPPI regressors of interest. We systematically varied whether to up-sample the seed amygdala timeseries to a sampling resolution of 10Hz (effective TR = 0.1s), and whether to apply no regularization versus an L1 penalty to the deconvolution solution parameters. We then compared within-scan similarity of gPPI regressors by computing product-moment correlations between all generated gPPI regressors for a given scan. We also computed equivalent correlations between gPPI regressors and the seed timeseries.

Within-person similarity of voxel-wise statistical maps for amygdala FC from gPPI pipelines

We asked whether, for each given scan, whether different pipelines yielded similar voxel-wise patterns of estimates in a within-scan analysis. To explore the degree to which scan-level statistical maps for gPPI contrasts were impacted by whether a deconvolution step was included in the pipeline, we computed similarity statistics for fear > baseline, neutral > baseline, and fear > neutral gPPI contrast t-statistic maps between the pipelines with versus without deconvolution. Similarities were calculated using product-moment correlations of 3d images (using AFNI 3ddot; https://afni.nimh.nih.gov/pub/dist/doc/program_help/3ddot.html) after transformation to MNI space at 2mm (isotropic) resolution, and across all voxels in the brain. In addition to whole-brain similarity, we also computed similarity statistics in the same way for all voxels within all four mPFC ROIs previously described.

Within-person similarity between gPPI & BSC amygdala FC

Given previous evidence that gPPI and BSC connectivity methods detect similar differences in connectivity between task contrasts (Di et al., 2020), we asked whether this was true for the fear > neutral faces contrast, specifically for amygdala FC. To accomplish this, we computed mean FC with the Harvard-Oxford bilateral amygdala for each ROI in the Harvard-Oxford cortical and subcortical atlases (https://neurovault.org/collections/262/). For gPPI, mean FC was computed as the mean t-statistic over all voxels in each ROI for the fear > neutral contrast. For BSC, mean FC was computed as the product-moment correlation between the average timeseries of the bilateral amygdala and each ROI. We thus compiled a vector representing amygdala FC with 62 ROIs (we removed all subcortical ROIs representing white matter, ventricles, or entire hemispheres of cerebral cortex) across the brain for gPPI (both with and without deconvolution) and BSC pipelines (both with and without global signal subtraction (GSS)). For each scan, we computed product-moment correlations between each of these vectors as a measure of the similarity of amygdala FC with the rest of the brain across pipelines. We also computed within-person similarity for BSC amygdala connectivity for pipelines with versus without global signal subtraction.

Associations between amygdala—mPFC circuitry & separation anxiety:

Amygdala—mPFC circuitry & separation anxiety: multiverse details (Table 2, Aim 5)

Amygdala reactivity measures: We used t-statistic estimates from the bilateral amygdala (both native and MNI space) for fear > baseline, neutral > baseline, and fear > neutral contrasts as the predictor of interest.

Amygdala reactivity slope measures: For analyses of change in amygdala reactivity over trials, we used the slope calculated across fear and neutral trials, respectively, as previously described. We used a bilateral amygdala mask (in MNI space) to define amygdala reactivity slopes, and included pipelines both with and without global signal correction.

Amygdala—mPFC FC measures: We used t-statistic estimates for gPPI between bilateral amygdala (in MNI space) and all four mPFC ROIs previously described, as well as BSC estimates for FC between the same regions. We included gPPI estimates both with and without a deconvolution step, and BSC estimates with and without global signal correction. We submitted estimates to group-level models for fear, neutral, and fear > neutral contrasts.

Separation anxiety outcomes: Analyses were run for each of three separation anxiety outcomes: scores from the RCADS separation anxiety subscale, and both raw and standardized t-scores from the SCARED separation anxiety subscale.

Group-level models: In all models, separation anxiety outcomes were modeled using multilevel linear regressions with crossed random effects for age and brain measure. All separation anxiety models were formulated to be robust to outliers (using Student's t likelihood functions), and we did not exclude previously analyzed scans from any analysis pipelines.

Longitudinal model syntax for amygdala—mPFC circuitry & separation anxiety

Longitudinal models for associations between amygdala—mPFC circuitry and separation anxiety included participant-varying slopes for the respective brain measure included as a predictor in each model, as well as for age.

Separation anxiety model

```
brm(separation_anxiety ~ brain_measure + age_centered + motion + (brain_measure +
age_centered|participant, chains = 4, cores = 4,
        family = 'student', prior = prior(gamma(4, 1), class = nu))
```

Estimating impacts of specific forked decision points on age-related change estimates

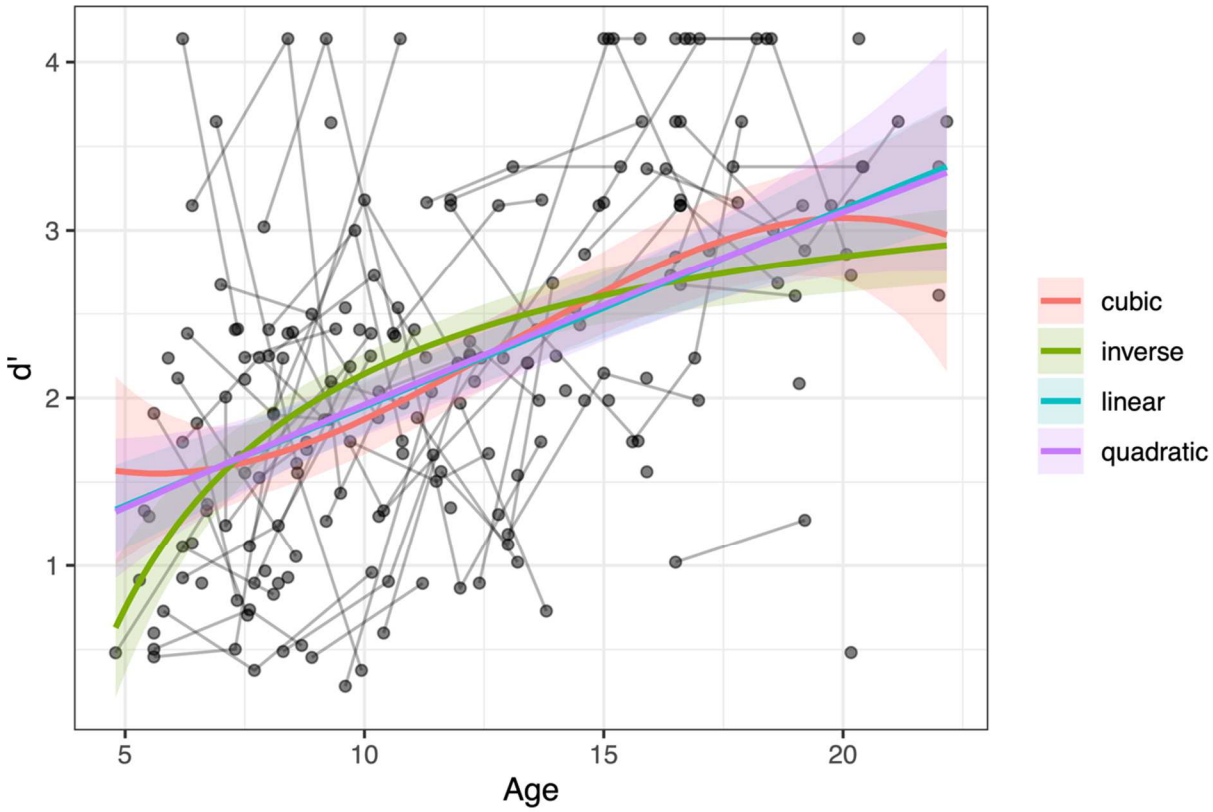
To explore impacts of different analytical decision points (or the impacts of 'taking different forks' in the path from beginning to end of the analysis) on age-related change estimates, we submitted point estimates for linear age-related change (posterior medians) from each model to separate Bayesian linear regression models. Models were fit using the rstanarm package (Gabry et al., 2019), and included each decision point (one-hot encoded if there were more than 2 options) as a binary predictor of the point estimates for age-related change. Following modeling, we plotted posterior distributions and 95% posterior intervals for each parameter, representing effects of each decision point conditional on all others. Example syntax of one model for amygdala reactivity decision points is below.

```
decision_point_model = stan_glm(data = sca_frame, estimate ~ tstat + quadratic +
                random_slopes + ctrl_scanner + ctrl_block + exclude_prev +
                amyg_right + amyg_left + amyg_high_sig + amyg_low_sig +
                native_space + motion_reg6 + motion_reg18 +
                hrf_2gamma + highpass + robust,
                cores = 4, chains = 4)
```

Supplemental Results

Behavior: Supplemental Results

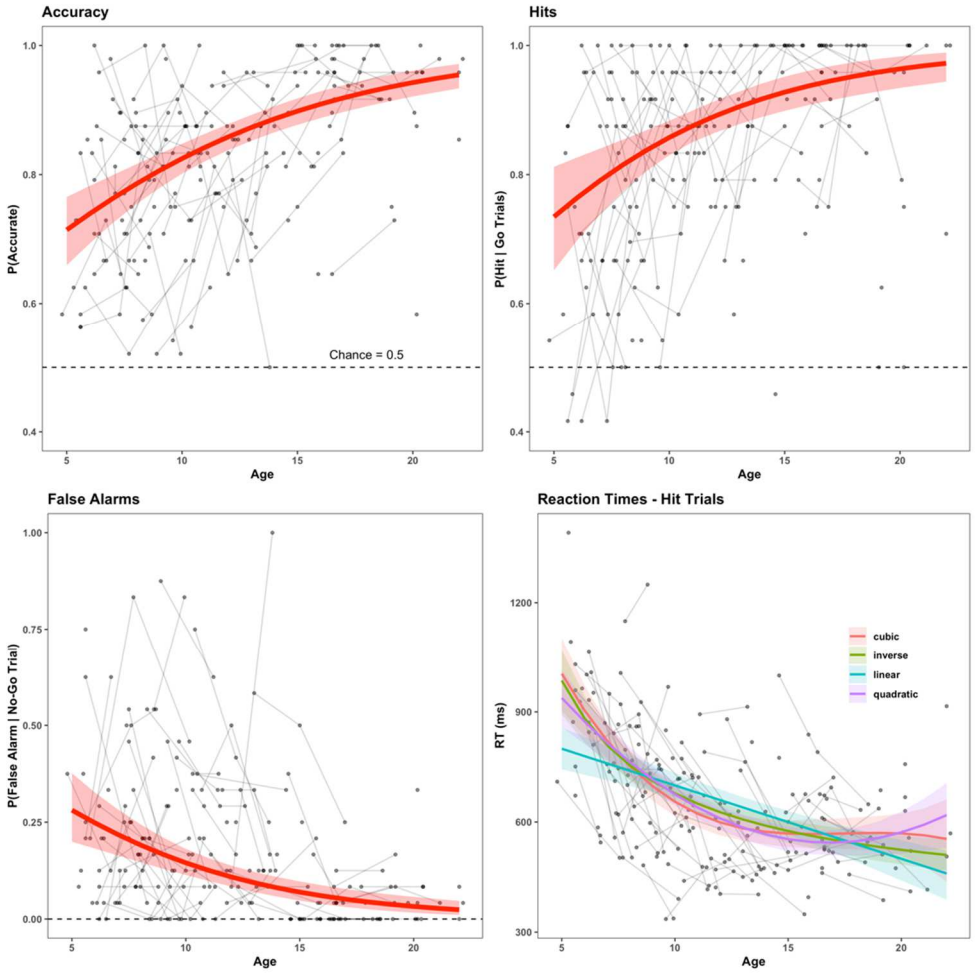
Behavioral Task Performance



Appendix A Figure 3: d' as a function of age.

d' for task performance as a function of age, modeled using linear, quadratic, cubic, and inverse age longitudinal models. Points display performance for one participant at one timepoint, black lines connect estimates from participants with multiple study visits, and colored lines with shaded area represent fitted model predictions and 95% posterior intervals.

We modeled task performance with d' (using a correction for extreme proportions; Hautus, 1995) as a function of linear, quadratic, cubic, and inverse age trends (Appendix A Figure 3). Models indicated age-related improvements in d' without notable quadratic or cubic change. The ICC for d' scores was estimated to be 0.31 (95% PI [0.04, 0.51]). We also modeled accuracy (probability of a correct response on any trial, with hits and correct rejections coded as 1, misses and false alarms coded as 0), hits, and false alarms using single-trial multilevel logistic regression models, and found similar age-related increases in task performance (Appendix A Figure 4). Even at the youngest ages, task performance was well above chance levels. We also modeled reaction time as a function of age using linear, quadratic, and cubic models. Along with general age-related decreases in reaction times to neutral faces ('hits'), quadratic and cubic models indicated that the most age-related change in average reaction times occurred at the younger end of the age range, between approximately 4-11 (Appendix A Figure 4, bottom right panel).

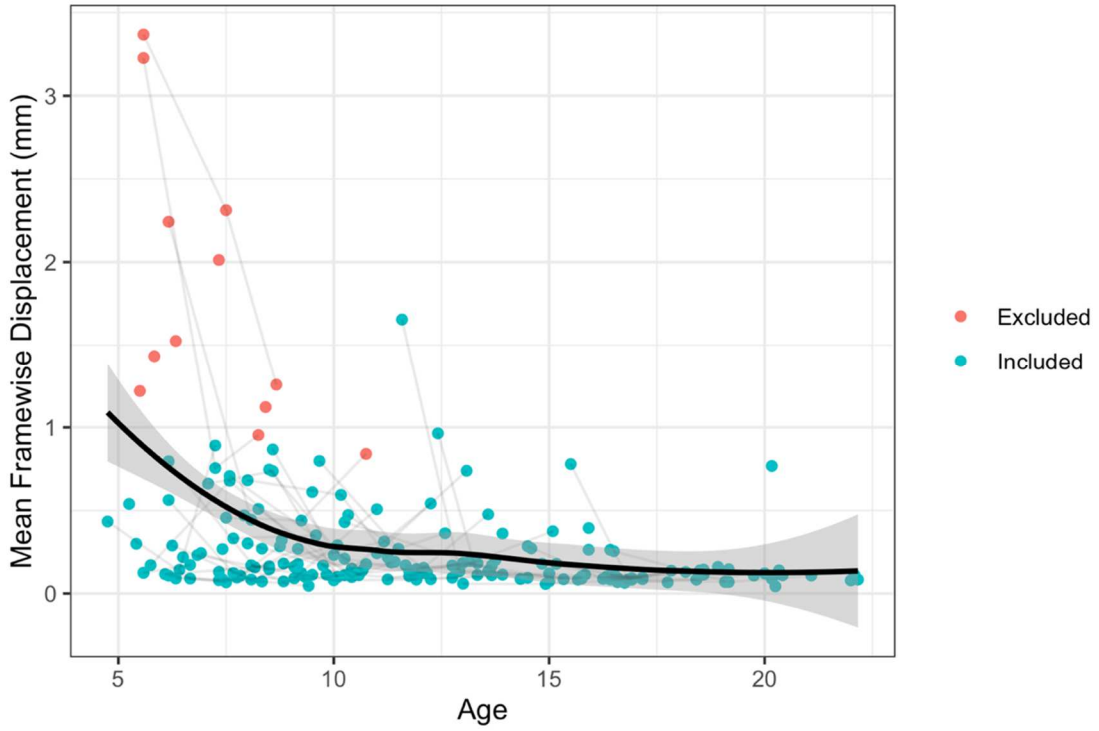


Appendix A Figure 4: Task performance metrics as a function of age

Age-related change in accuracy (top left), hits (top right), false alarms (bottom left), and reaction times during go trials (bottom right). Points display summarized performance for one participant at one timepoint (e.g. the proportion accurate across trials during that run), black lines connect performance summaries from participants with multiple study visits, and colored lines with shaded area represent fitted trial-by-trial model predictions and 95% posterior intervals.

Head Motion

Head motion, as measured by mean framewise displacement estimates derived from MCFLIRT (Jenkinson et al., 2002) decreased with age ($\hat{\beta}$=-0.04, 95% PI [-0.05, -0.02]) among all participants with available task fMRI data (Appendix A Figure 5). Among only included participants with <= 40 TRs with FD < 0.9mm, mean framewise displacement also decreased with age, although somewhat less strongly ($\hat{\beta}$=-0.02, 95% PI [-0.02, -0.01]).

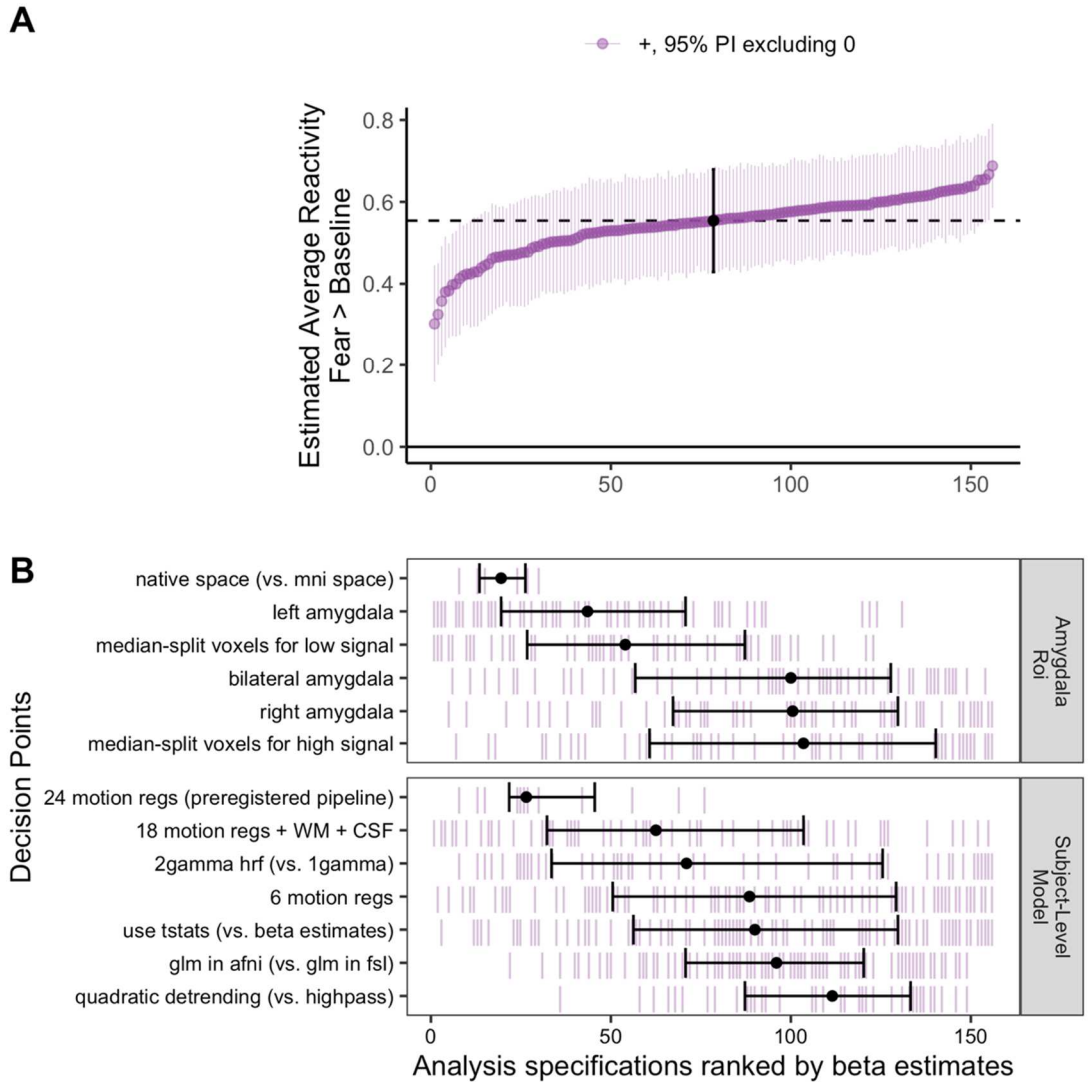


Appendix A Figure 5: Mean framewise displacement as a function of age (years)

Amygdala Reactivity: Supplemental Results

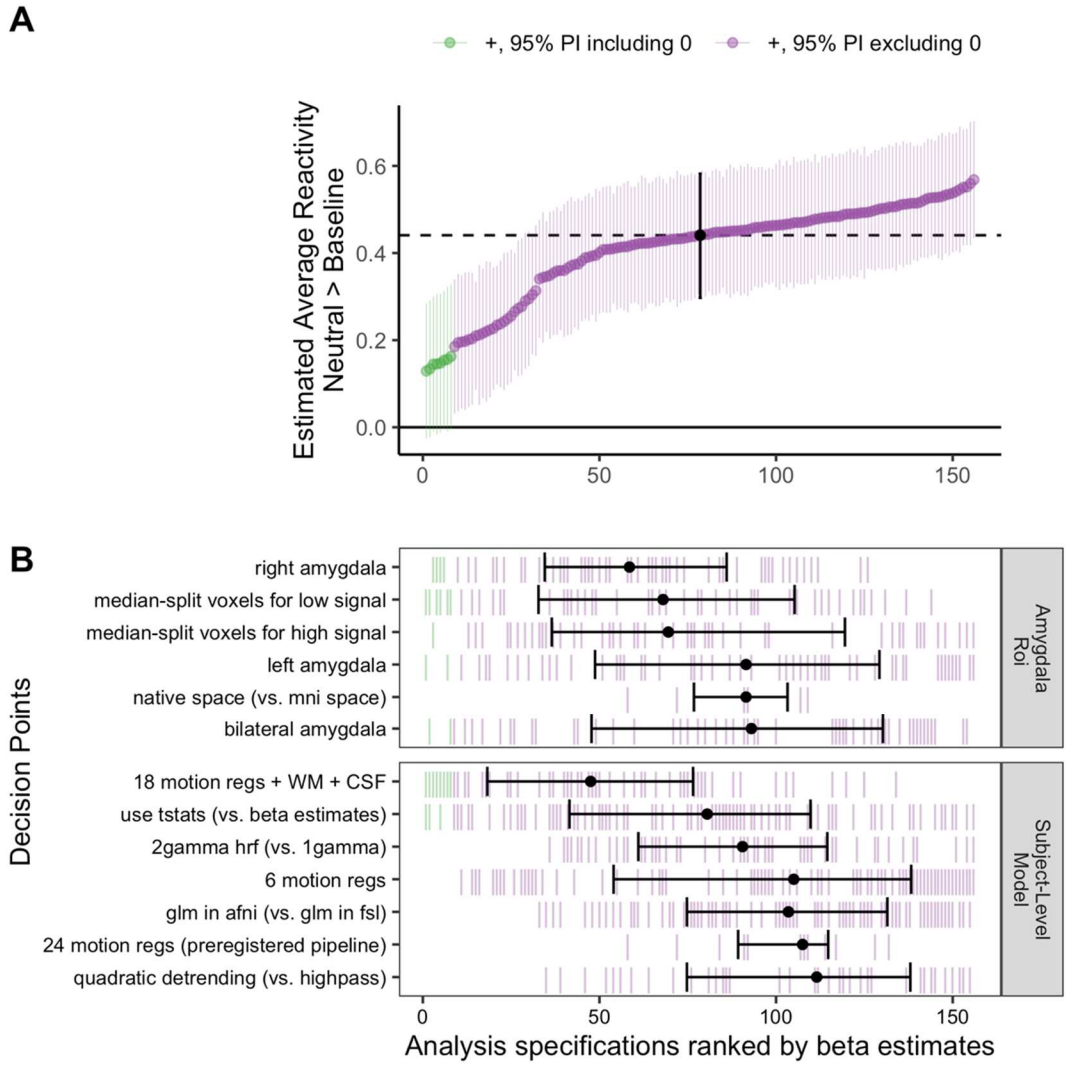
Group Mean Amygdala Reactivity

We separate specification curves to estimate group averages in amygdala reactivity for the fear > baseline, neutral > baseline, and fear > neutral contrasts. To preserve computational resources, all models for mean amygdala reactivity were fit using the lme4 R package with intercepts allowed to vary by participant. All specifications for the fear > baseline contrast (Appendix A Figure 6) and most specifications for the neutral > baseline contrast (Appendix A Figure 7) resulted in positive amygdala reactivity with a confidence interval distinct from zero, indicating robust average amygdala reactivity to faces of both emotions. In addition, most specifications found higher amygdala reactivity for fear compared to neutral faces (fear > neutral contrast, Appendix A Figure 8). Higher average amygdala reactivity for fear > neutral faces may have been due to either the face emotions or the fact that participants were instructed to press for neutral faces and withhold button press for fear faces.



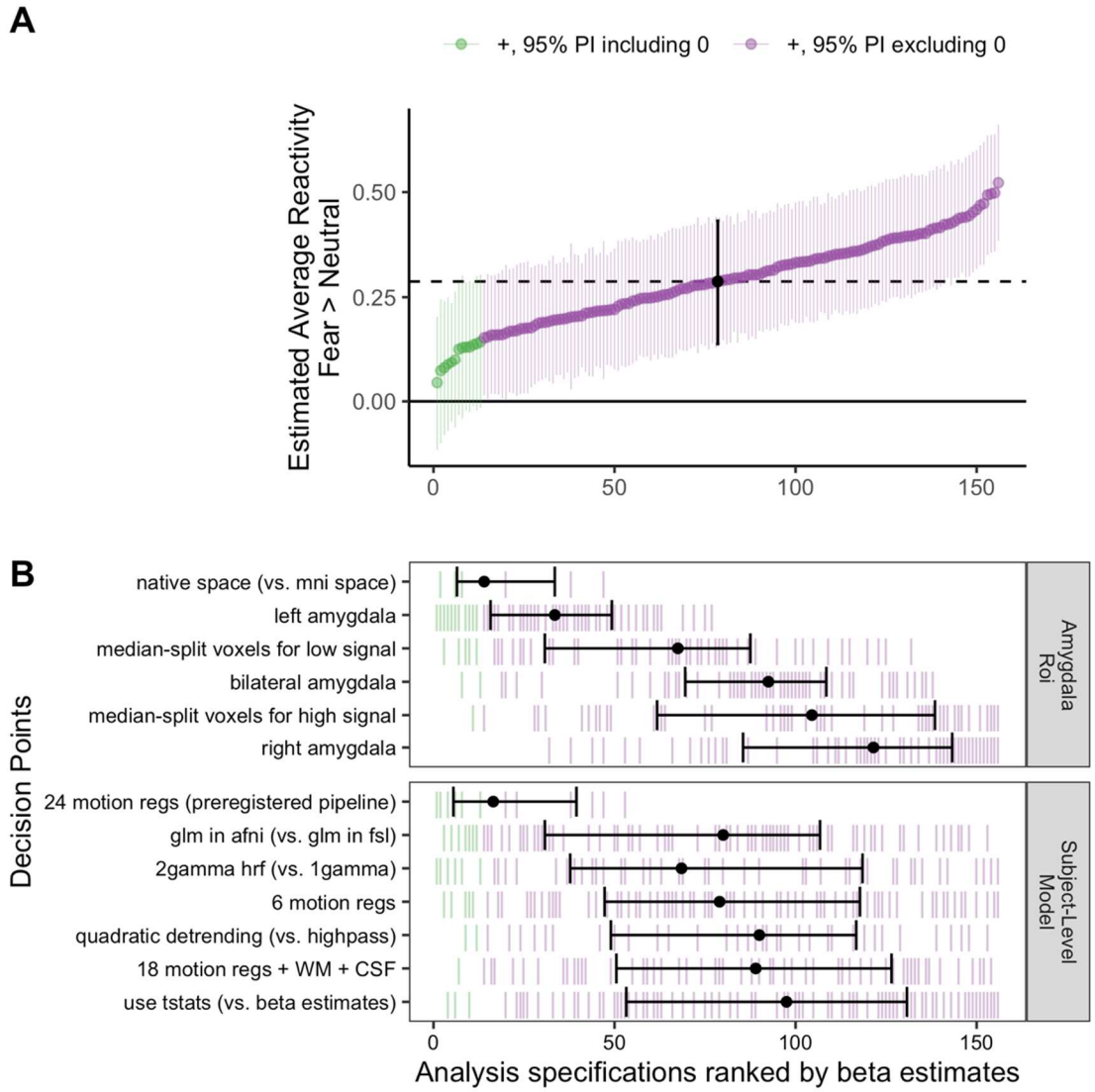
Appendix A Figure 6: Specification curve for mean fear > baseline amygdala reactivity

A: Points represent estimated mean amygdala reactivity for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.



Appendix A Figure 7: Specification curve for mean neutral > baseline amygdala reactivity

A: Points represent estimated mean amygdala reactivity for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.

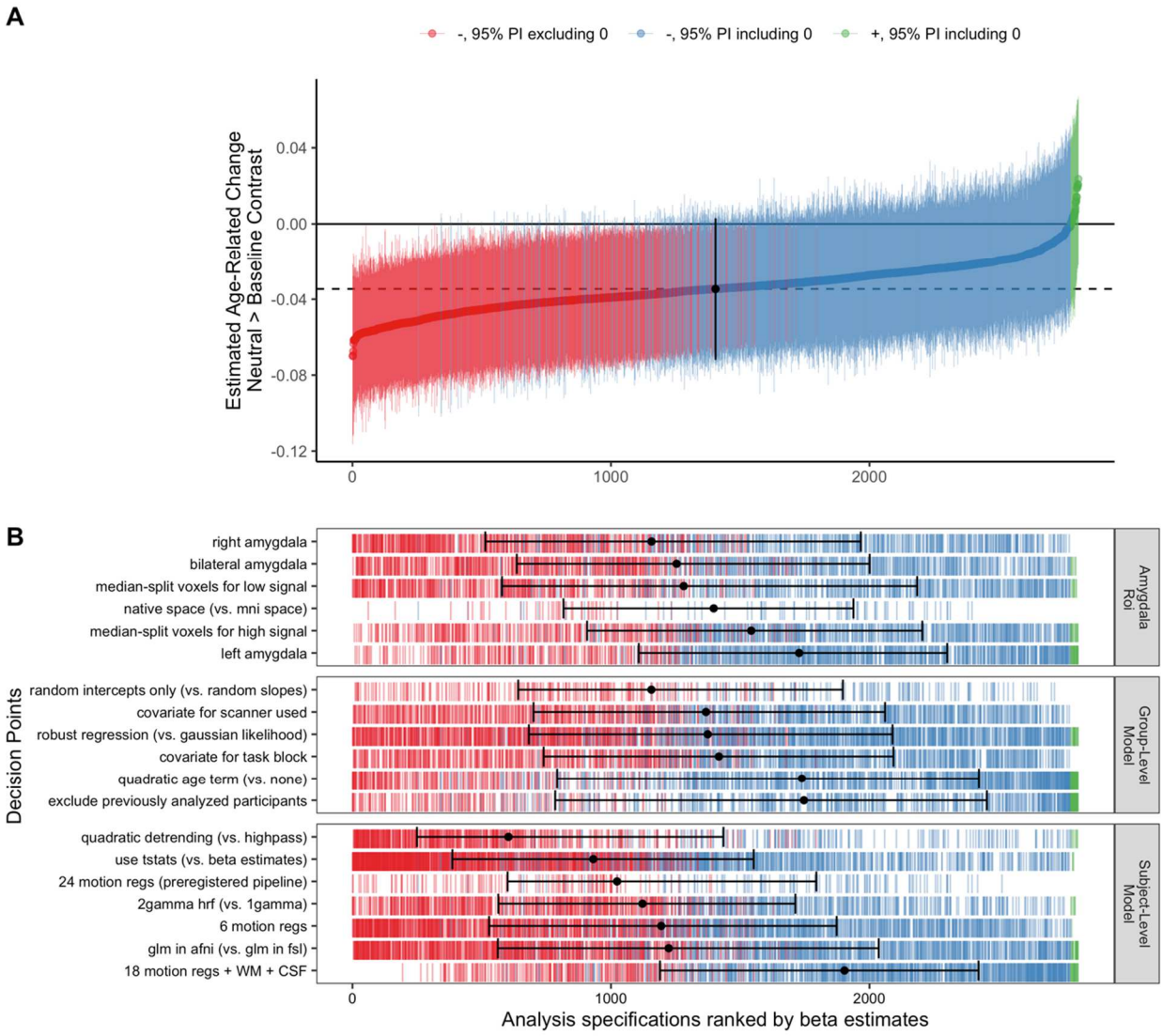


Appendix A Figure 8: Specification curve for mean fear > neutral amygdala reactivity

A: Points represent estimated mean amygdala reactivity for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.

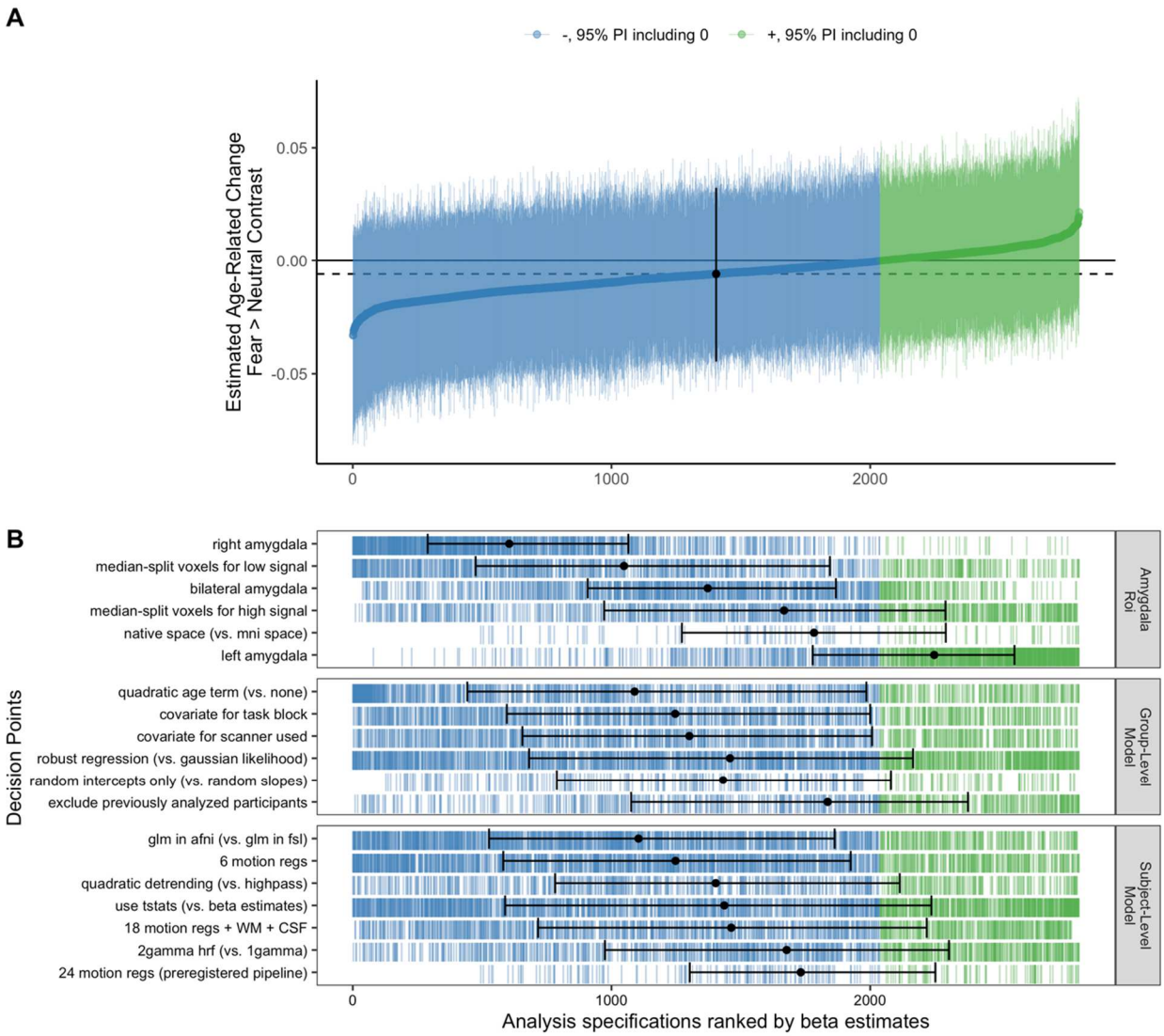
Multiverse analyses of age-related changes in amygdala reactivity

In addition to constructing specification curves for age-related change in amygdala reactivity for the fear > baseline contrast as reported in the main manuscript (see Figure 2), we constructed parallel specification curves for the neutral > baseline and fear > neutral contrasts. Generally, age-related change findings for the neutral > baseline contrast were similar to, but slightly weaker than, fear > baseline: while 98.9% of specifications found negative age-related change, only 42.4% of specifications estimated negative age-related change distinguishable from 0 (see Appendix A Figure 9). No pipelines found age-related change in fear > neutral amygdala reactivity distinguishable from 0 (see Appendix A Figure 10).



Appendix A Figure 9: Specification curve for age-related change in neutral > baseline amygdala reactivity

A: Points represent estimated linear age-related change in amygdala reactivity for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.



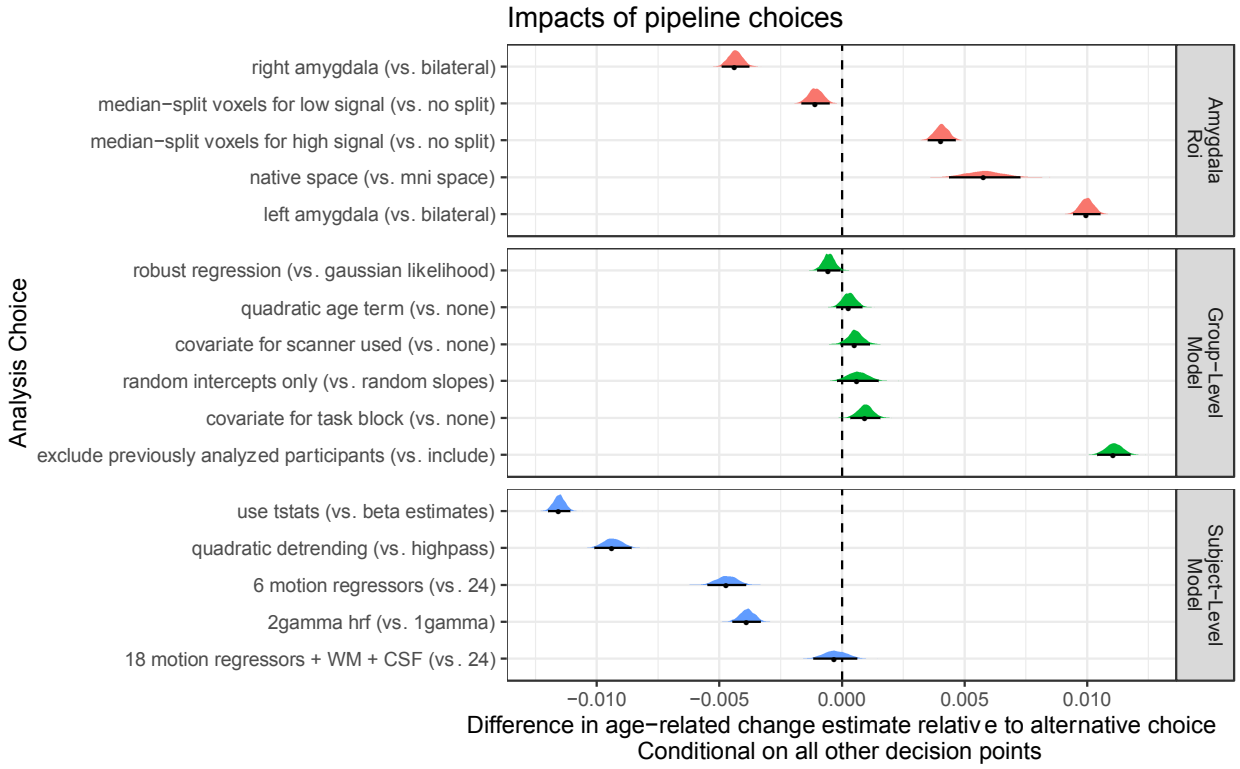
Appendix A Figure 10: Specification curve for age-related change in fear > neutral amygdala reactivity

A: Points represent estimated linear age-related change in amygdala reactivity for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-

axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.

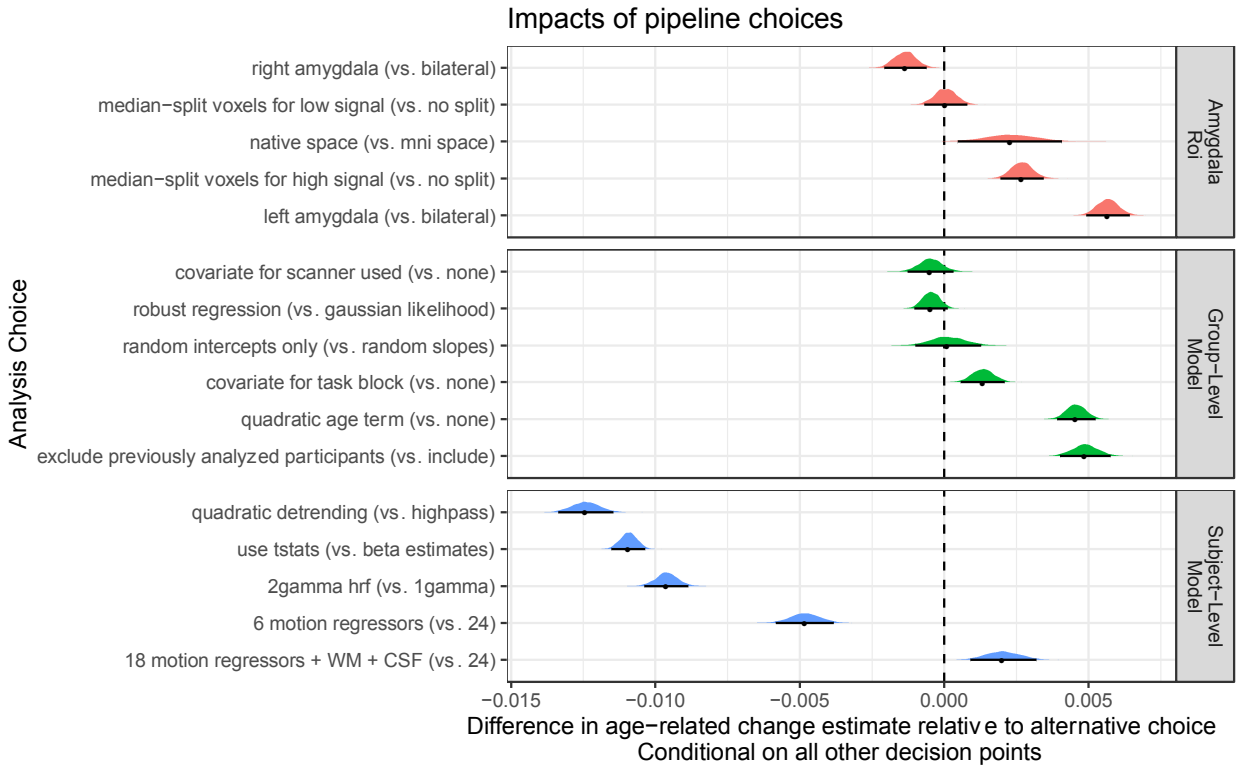
Impacts of pipeline choices on age-related change estimates for amygdala reactivity

While the goals of the present study were not to precisely quantify impacts of specific pipeline decision points, we explored impacts of all decision points on estimates of age-related changes in amygdala reactivity for each contrast. Specifically for the fear > baseline contrast (see Appendix A Figure 11), age-related change was somewhat stronger (more negative) for specifications using a right amygdala region compared to a bilateral region, and weaker for specifications using a left amygdala region. Most notably, specifications excluding the 42 participants previously studied (Gee et al., 2013) found weaker age-related change.



Appendix A Figure 11: Fork impacts on age-related change for fear > baseline amygdala reactivity

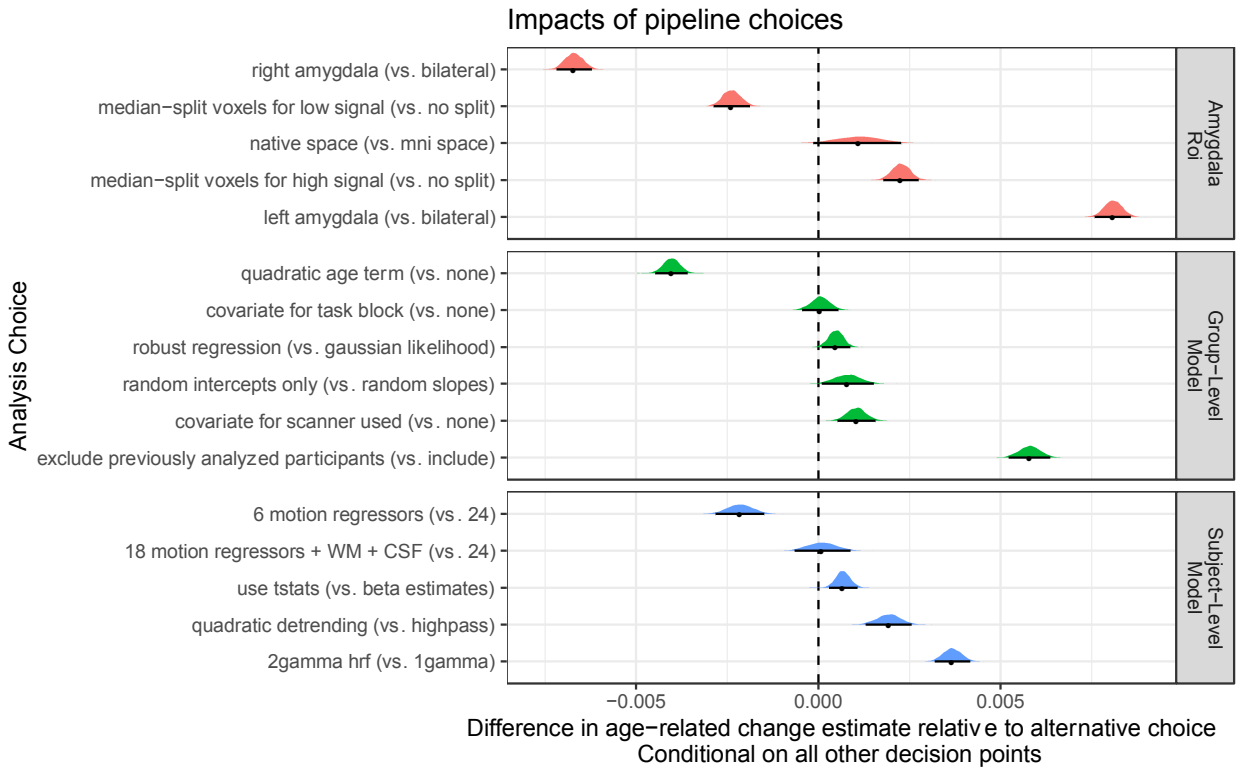
Posterior distributions and 95% posterior intervals are shown, representing the average difference in linear age-related change estimates relative to the alternative choice. As most specifications find negative age-related change, negative values indicate more strongly negative change, and positive values indicate weaker change.



Appendix A Figure 12: Fork impacts on age-related change for neutral > baseline amygdala reactivity

Posterior distributions and 95% posterior intervals are shown, representing the average difference in linear age-related change estimates relative to the alternative choice. As most

specifications find negative age-related change, negative values indicate more strongly negative change, and positive values indicate weaker change.



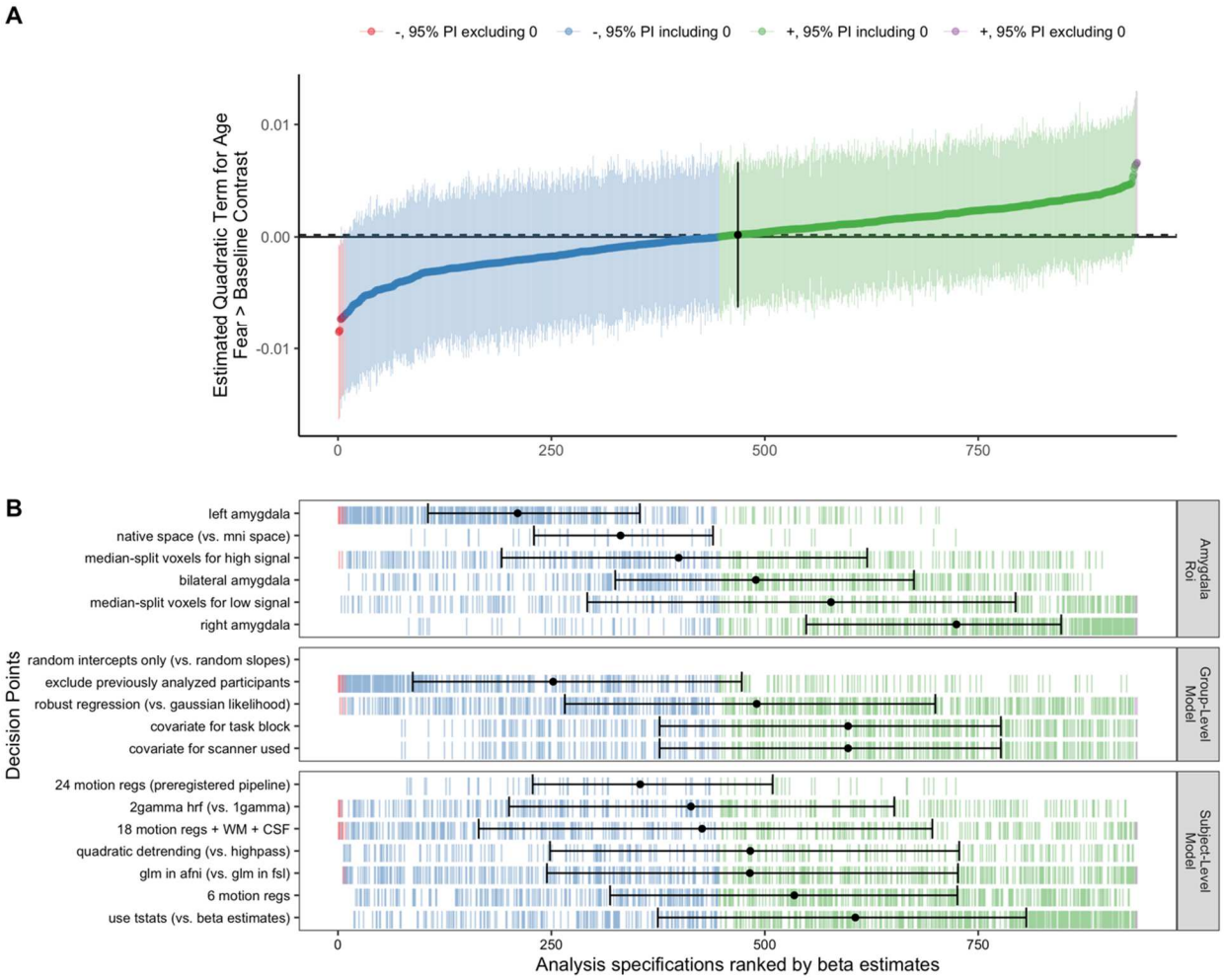
Appendix A Figure 13: Fork impacts on age-related change for fear > neutral amygdala reactivity

Posterior distributions and 95% posterior intervals are shown, representing the average difference in linear age-related change estimates relative to the alternative choice. Negative values indicate more negative change, and positive values indicate more positive change.

Nonlinear Changes in Amygdala Reactivity

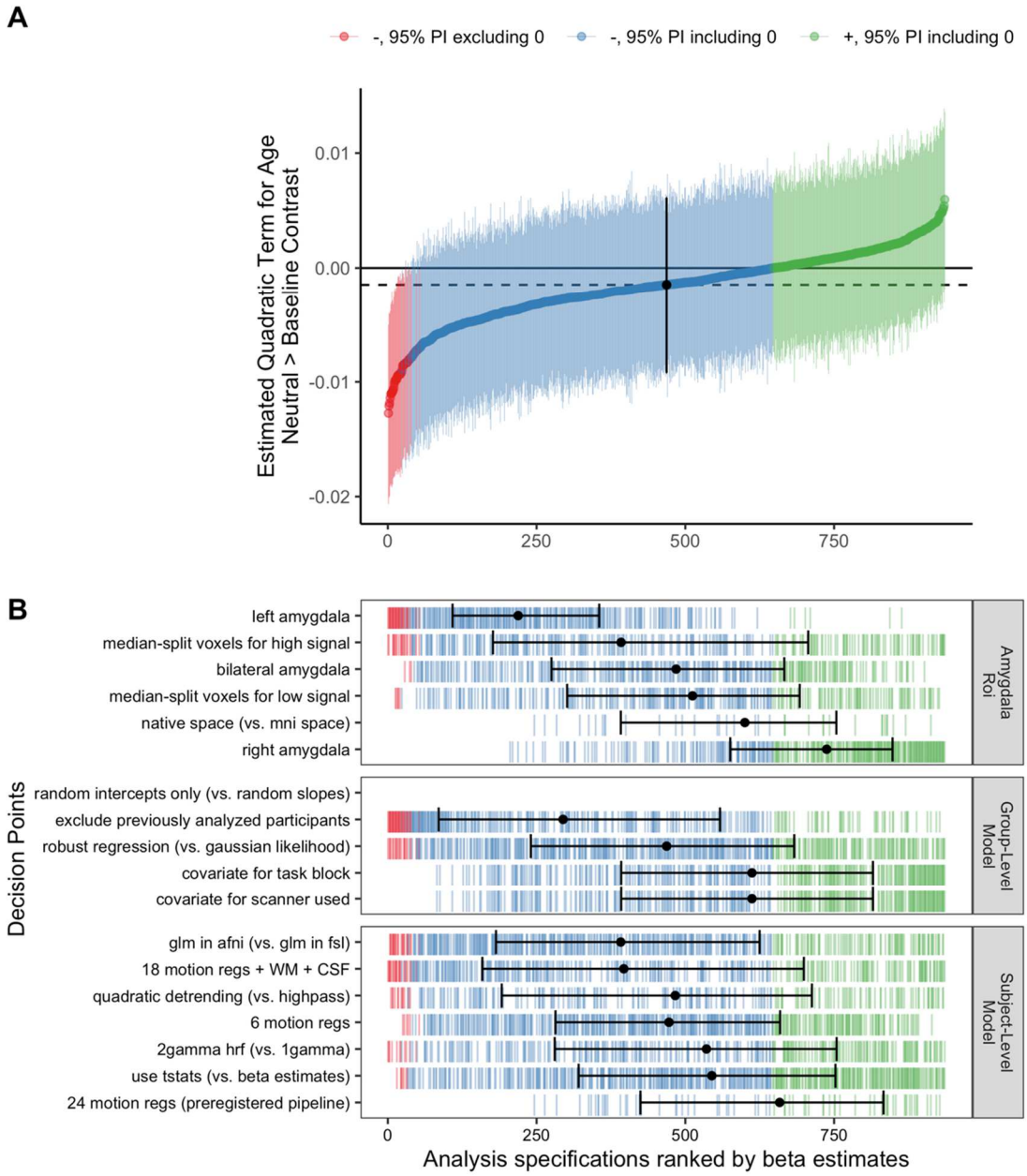
In addition to linear age-related change, we also included quadratic age terms in some model specifications for somewhat more flexible modeling of longitudinal trajectories of amygdala reactivity. We constructed specification curves for quadratic age-related change parameters across all models including quadratic terms for the fear > baseline (Appendix A Figure 14), neutral > baseline (Appendix A Figure 15), and fear > neutral (Appendix A Figure 16) contrasts. Across all contrasts, very few specifications estimated quadratic terms distinguishable from 0 under a 95% posterior interval. Further, quadratic fits were varied in sign for each contrast, such that some quadratic models estimated developmental 'peaks' while others estimated 'troughs' in amygdala reactivity. When model predictions from separate specifications were plotted as individual 'spaghetti', there was not clear consensus in quadratic trajectories for any contrast (as there was for linear change for the fear > baseline and neutral > baseline contrasts, see Appendix A Figure 17). Thus, while the current study may not have been adequately powered to estimate quadratic age-related change, we did not find consistent evidence for either peaks or troughs in amygdala reactivity between ages 4-22.

Because inverse age models may be particularly useful in characterizing rapid change early in development, we also fit such models such that amygdala reactivity was estimated as a function of 1/age. Models were fit using maximum likelihood, and fits indicated age-related decreases in the fear > baseline and neutral > baseline contrasts (Appendix A Figure 18). However, such inverse age models used here can characterize rapid change earlier in development but not later in development by design (i.e. they are formulated to capture deceleration in adolescence and stabilization in young adulthood; Luna et al., 2021).



Appendix A Figure 14: Spec. curve for quadratic age-related change in fear > baseline amygdala reactivity

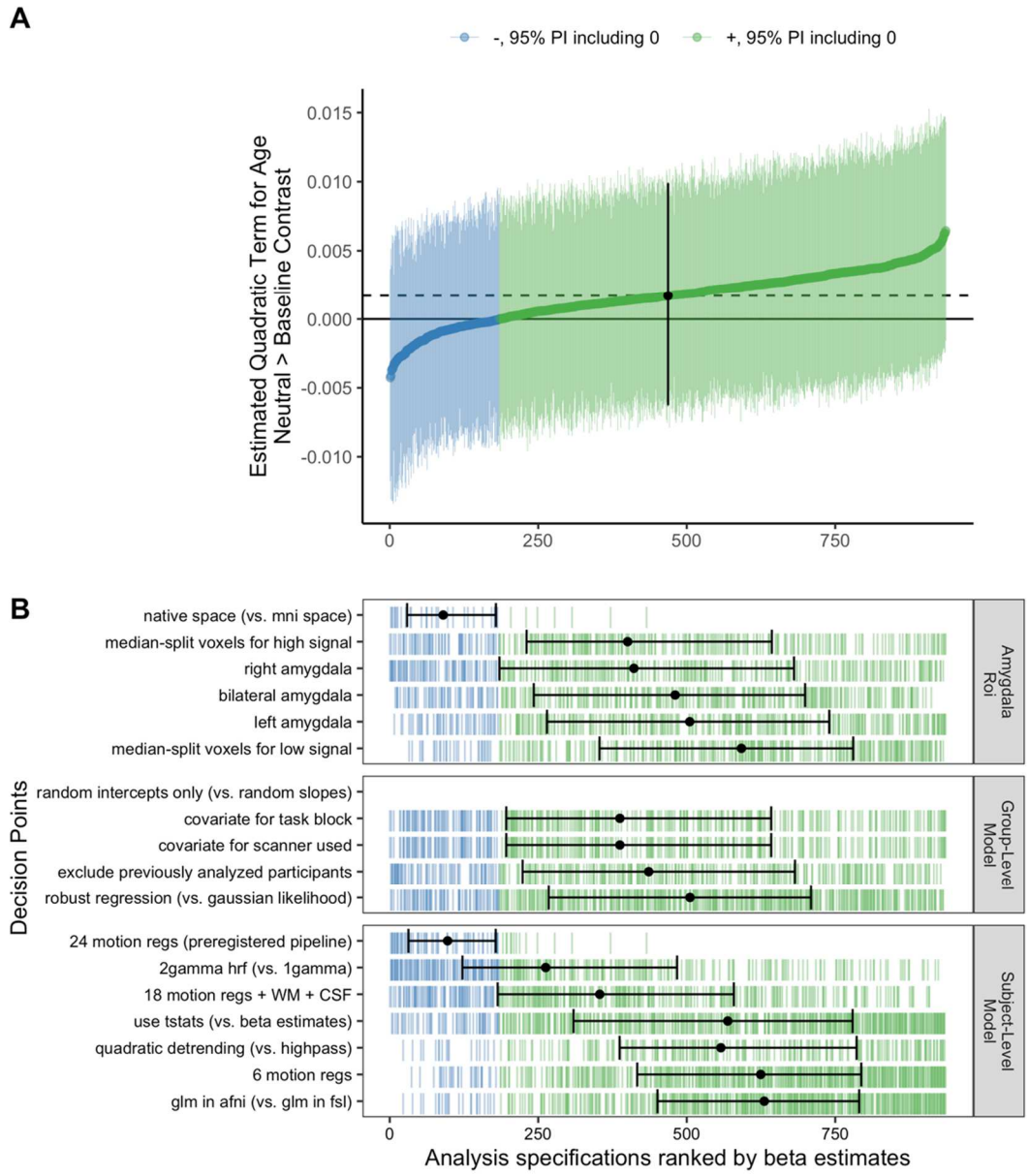
A: Points represent estimated quadratic age-related change in amygdala reactivity for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.



Appendix A Figure 15: Spec. curve for quadratic age-related change in neutral > baseline amygdala reactivity

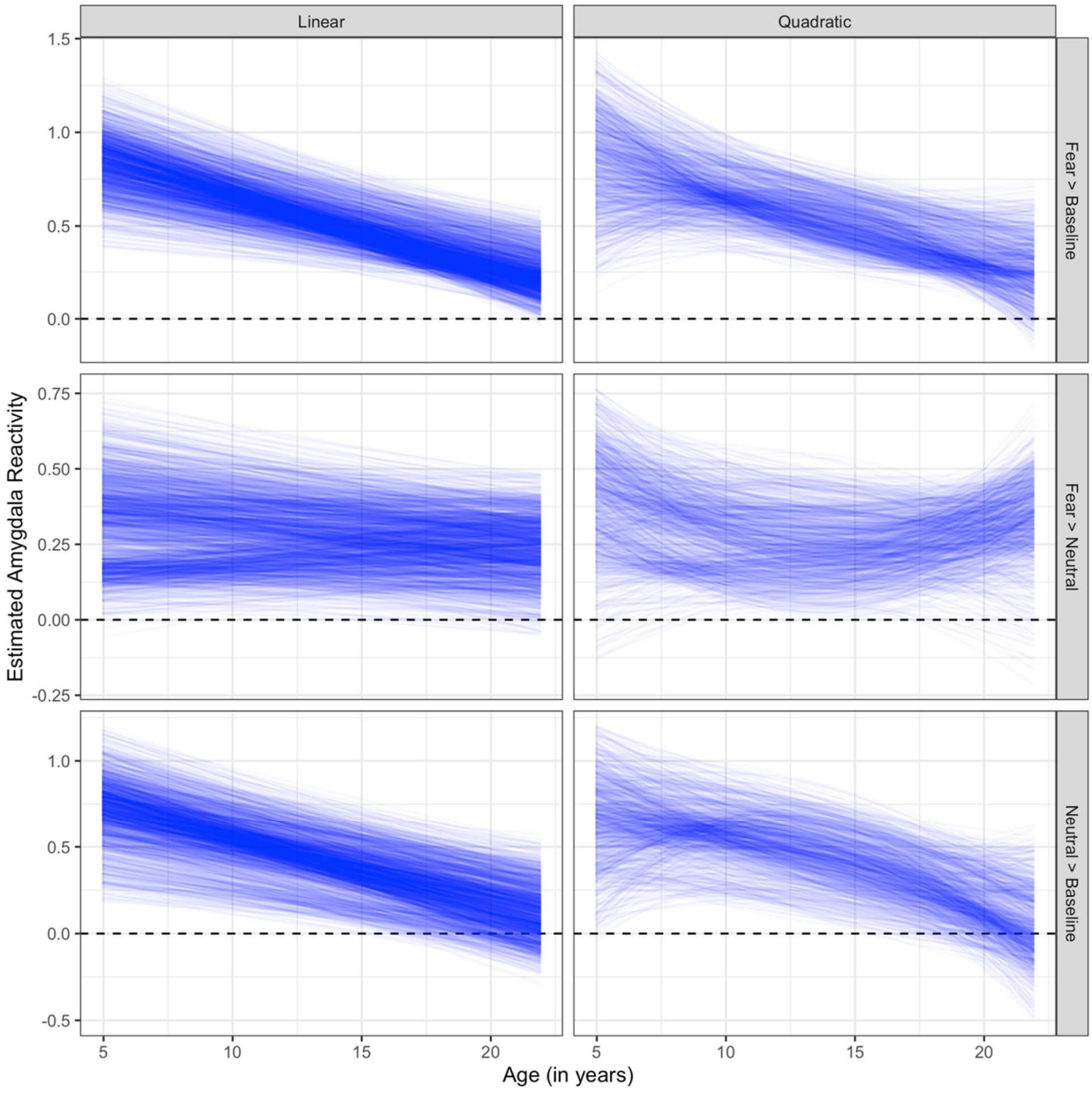
A: Points represent estimated quadratic age-related change in amygdala reactivity for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-

axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.

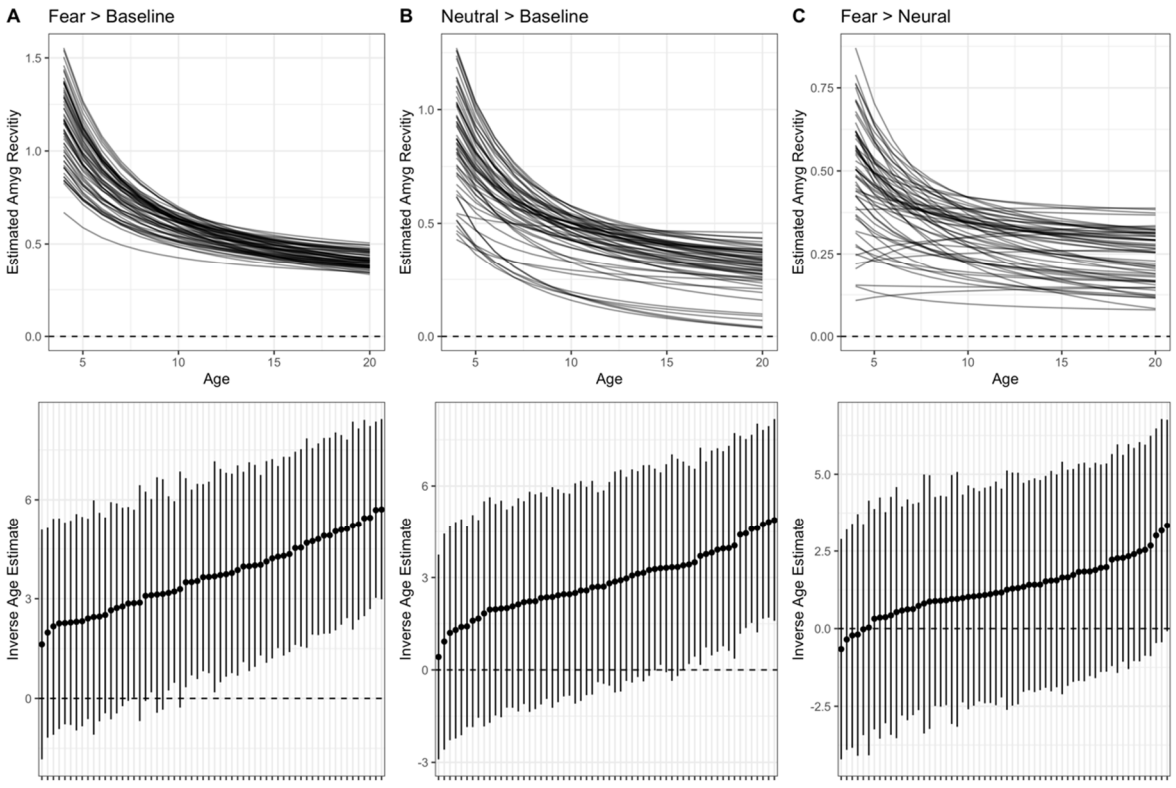


Appendix A Figure 16: Spec. curve for quadratic age-related change in fear > neutral amygdala reactivity

A: Points represent estimated quadratic age-related change in amygdala reactivity for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.



Appendix A Figure 17: Predictions of age-related change across linear and quadratic model specifications. Each blue line represents 1 specification, with age on the x-axis and estimated amygdala reactivity on the y-axis. For quadratic models (right panel), some specifications found convex change while other indicated concave change.



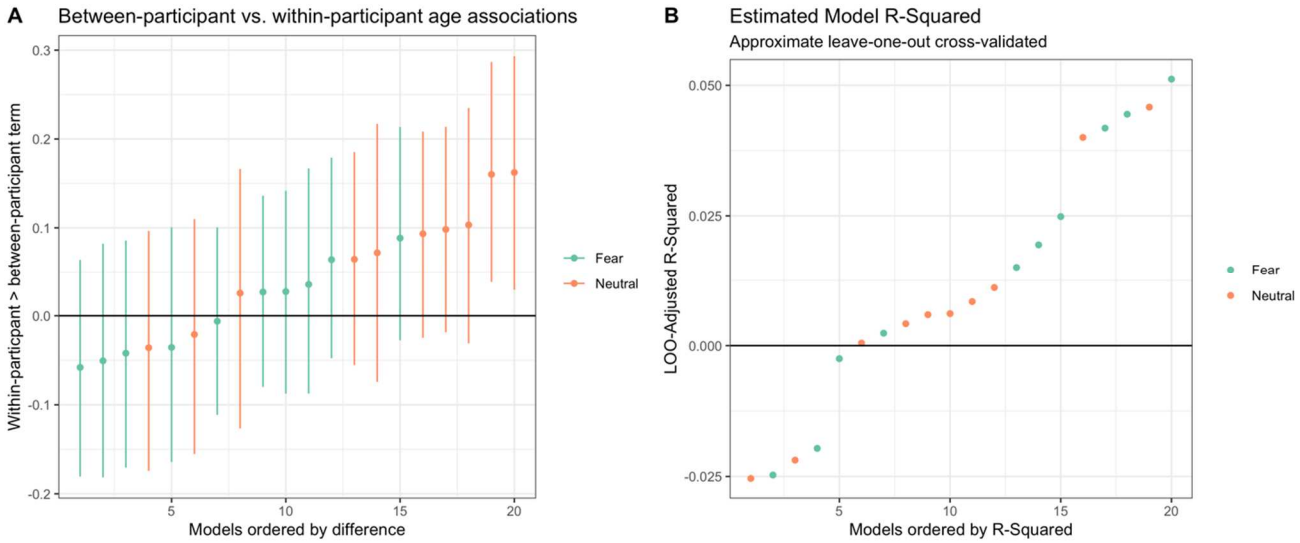
Appendix A Figure 18: Inverse age models for amygdala reactivity

Spaghetti plots for fitted model predictions (top) and specification curves (bottom) for inverse age estimates for models of amygdala reactivity. Positive parameter fits for inverse age indicate

decrease in amygdala reactivity as a function of age. Panels represent the fear > baseline contrast (A, left), neutral > baseline contrast (B, center), and the fear > neutral contrast (C, right)

Between-participant age associations versus within-participant age-related change

We constructed a smaller specification curve of models parametrized to differentiate between-participant age associations from within-participant age-related changes in amygdala reactivity. While only between-participant terms indicated consistent associations with age, we also examined, for each specification, whether between-participant and within-participant age terms *differed* from one another. We estimated the differences between such terms through calculating the distribution of paired differences in posterior draws between the two terms, and summarizing using the median and 95% quantiles (to construct a 95% posterior interval). Overall, we found that despite the higher estimation precision for between-participant age associations reported in the main text (see Figure 2), within-participant and between-participant estimates did not consistently differ within most models for fear > baseline or neutral > baseline amygdala reactivity (Appendix A Figure 19A). In addition, R^2 metrics calculated with approximate leave-one-out cross-validation were close to 0 for all models, indicating predictive performance close no better than chance (Appendix A Figure 19B).



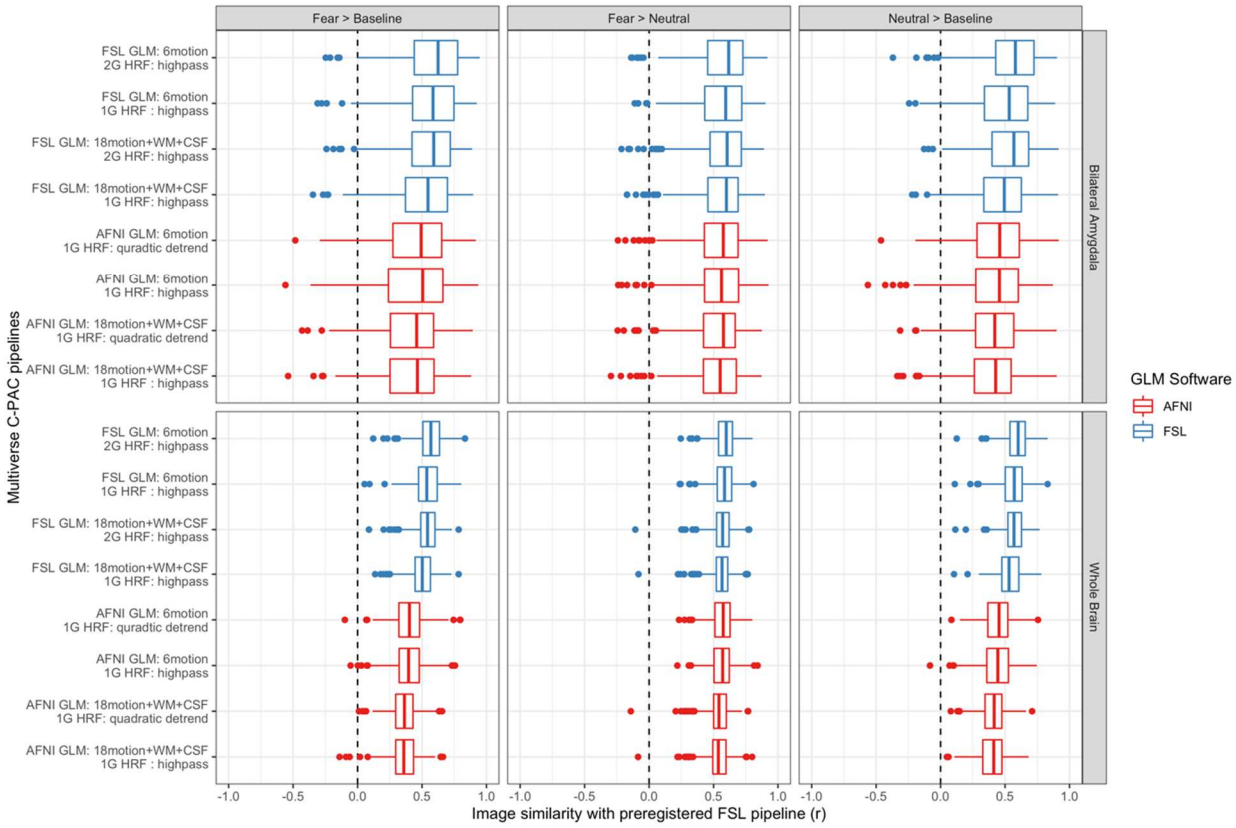
Appendix A Figure 19: Differences between within-participant and between-participant terms for age-related associations with amygdala reactivity

A. Estimated differences between within-participant and between-participant terms for age-related associations with amygdala reactivity. B. Approximated leave-one-out cross-validated R^2 scores for each specification separately parametrizing within-participant and between-participant terms. R^2 values below 0 indicate cross-validated performance poorer than that expected under the "null" model.

Within-person similarity of voxel-wise amygdala reactivity statistical maps across forks

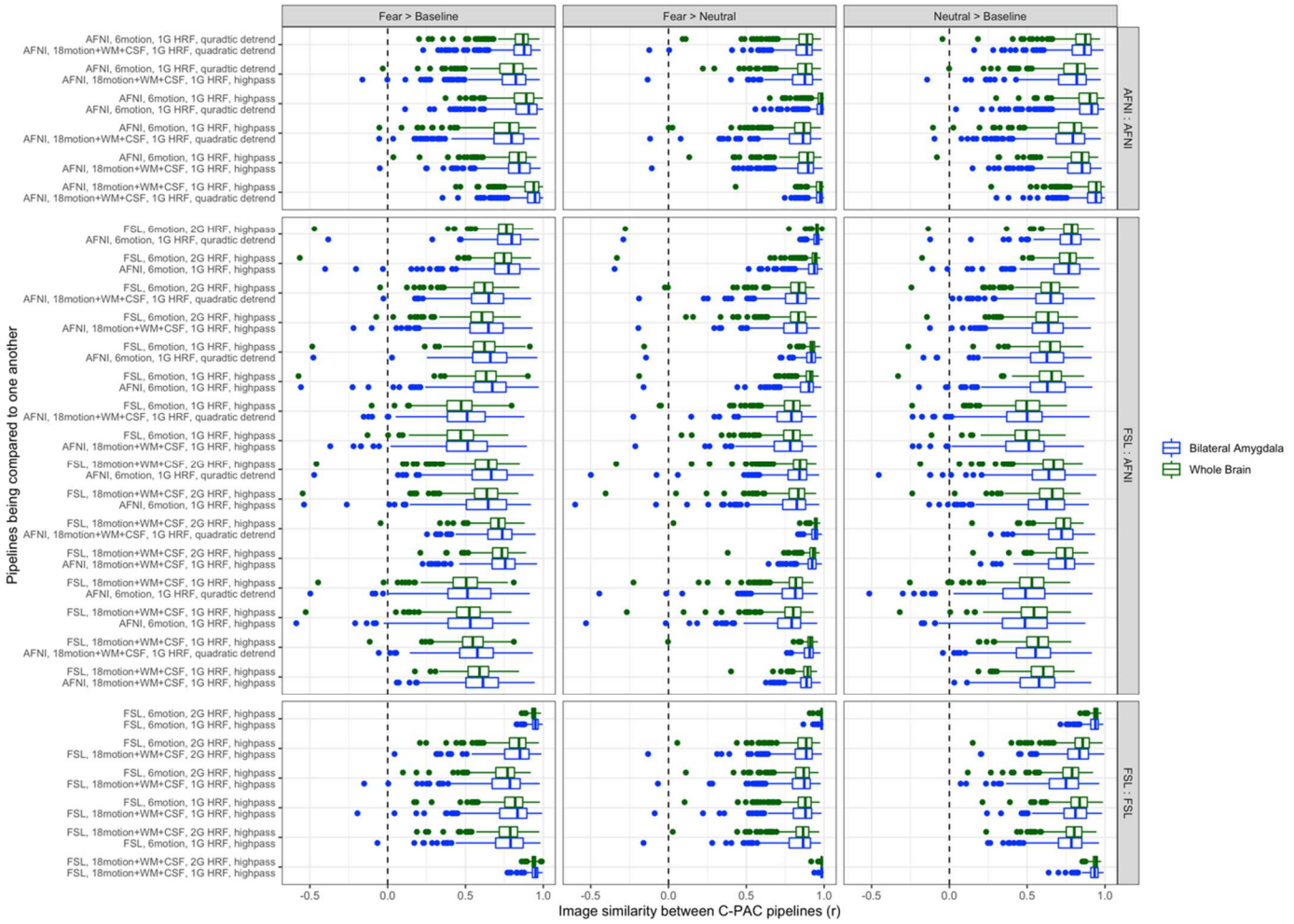
We sought to understand whether different preprocessing pipelines, for each given scan, yielded similar voxel-wise patterns of estimates in a within-scan analysis. To understand which preprocessing steps influenced reactivity estimates most, we computed the voxel-wise similarity of statistical maps (t-statistics) for the *same scans* across preprocessing specifications. Thus, for

each pair of pipelines, we computed 1 similarity value for each scan for the whole brain, and 1 similarity value for each scan for the bilateral amygdala. While similarity (product-moment correlation across all brain voxels) was positive for most comparisons between the preregistered pipeline (all FSL) and pipelines using C-PAC preprocessing, similarity did vary across scans such that for any comparison, some scans were highly different across specifications (i.e. near-zero or negative correlation values). In general, statistical maps for both the whole brain and the bilateral amygdala from the preregistered FSL pipeline were more similar to pipelines using C-PAC preprocessing + FSL GLMs compared to pipelines using C-PAC preprocessing + AFNI GLMs (see Appendix A Figure 20). In addition, within specifications with C-PAC preprocessing, similarity across pipelines was somewhat higher, especially within pipelines using the same GLM software or nuisance regressors (see Appendix A Figure 21). Relatively higher similarity among of specifications with C-PAC preprocessing was likely to do the common registration shared by all such pipelines (as opposed to differing registrations from the preregistered FSL pipeline).



Appendix A Figure 20. Voxelwise image similarity for amygdala reactivity contrasts between the preregistered FSL pipeline and C-PAC pipelines.

The x-axis indicates product-moment correlations across all brain voxels (bottom panel) and bilateral amygdala voxels (top) for the preregistered FSL pipeline with the pipelines indicated on the y-axis. Similarity is shown for the fear > baseline (left), fear > neutral (middle), and neutral > baseline (right) contrasts.

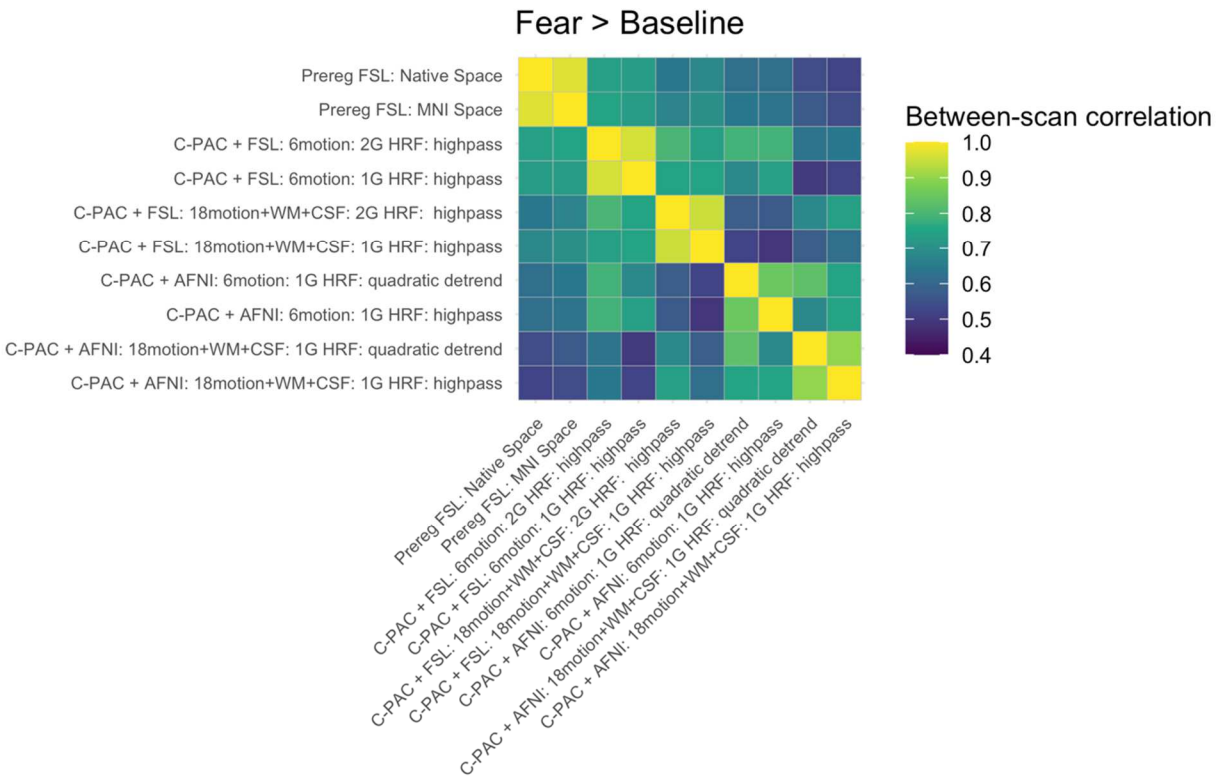


Appendix A Figure 21: Voxelwise similarity for amygdala reactivity contrasts between all C-PAC pipelines.

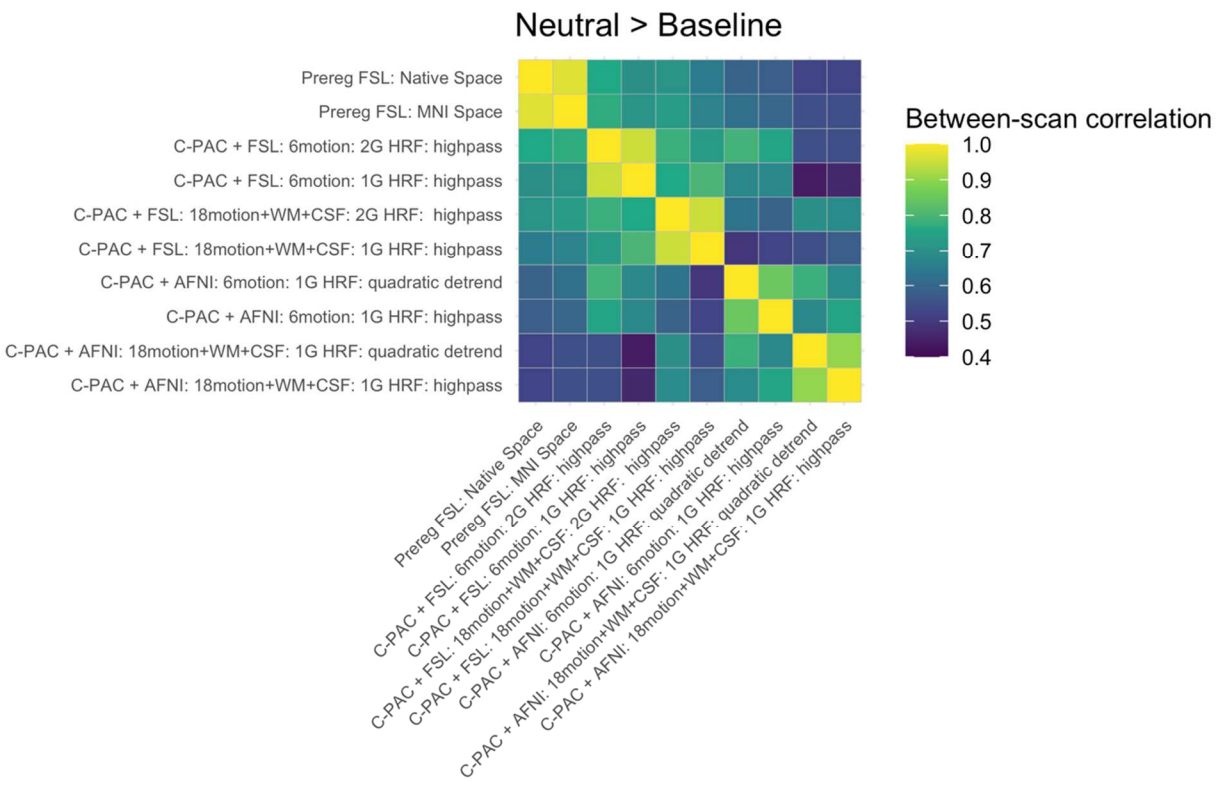
The x-axis indicates product-moment correlations across all brain voxels (green) and bilateral amygdala voxels (blue) within each scan for all pairwise comparisons of C-PAC pipelines indicated on the y-axis. Similarity is shown for the fear > baseline (left), fear > neutral (middle), and neutral > baseline (right) contrasts. Pipeline comparisons are organized into comparisons between two pipelines with AFNI GLMs (top), one pipeline with an AFNI GLM and the other with FSL (middle), and two pipelines with FSL GLMs (bottom).

Between-scan correlations of amygdala reactivity estimates across specifications

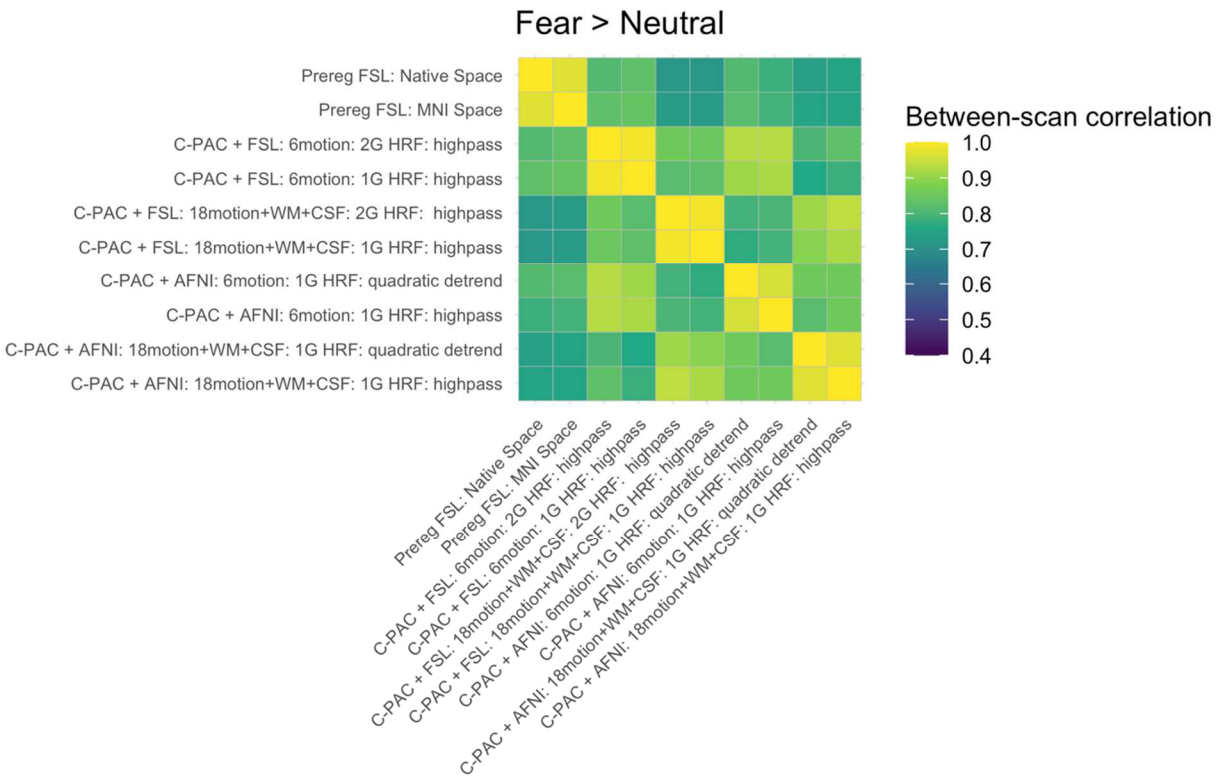
To ask whether relative relationships between scans for amygdala reactivity were preserved across preprocessing specifications, we computed correlations between scan-level estimates of mean amygdala reactivity across pairs of pipelines (for bilateral amygdala estimates, t-statistics) for each contrast. For each pair of preprocessing pipelines, we computed the rank-order correlation between vectors of bilateral amygdala reactivity estimates (1 datapoint per scan per pipeline). While correlations were all positive and mostly strong (r >= .7) for most pairs of pipelines for the fear > baseline (Appendix A Figure 22) and neutral > baseline (Appendix A Figure 23) contrasts, correlations between the most disparate preprocessing pipelines were often much weaker (e.g. from 0.4-0.6). Thus, the between-scan relationships between amygdala reactivity estimates were only somewhat weakly preserved across such pipelines. On the other hand, for the fear > neutral contrast, all pairwise comparisons of pipelines yielded higher (r >=.7) between-scan correlation values (Appendix A Figure 24).



Appendix A Figure 22: Between scan correlations for amygdala reactivity pipelines for fear > baseline.



Appendix A Figure 23: Between-scan correlations for amygdala reactivity pipelines for neutral > baseline.

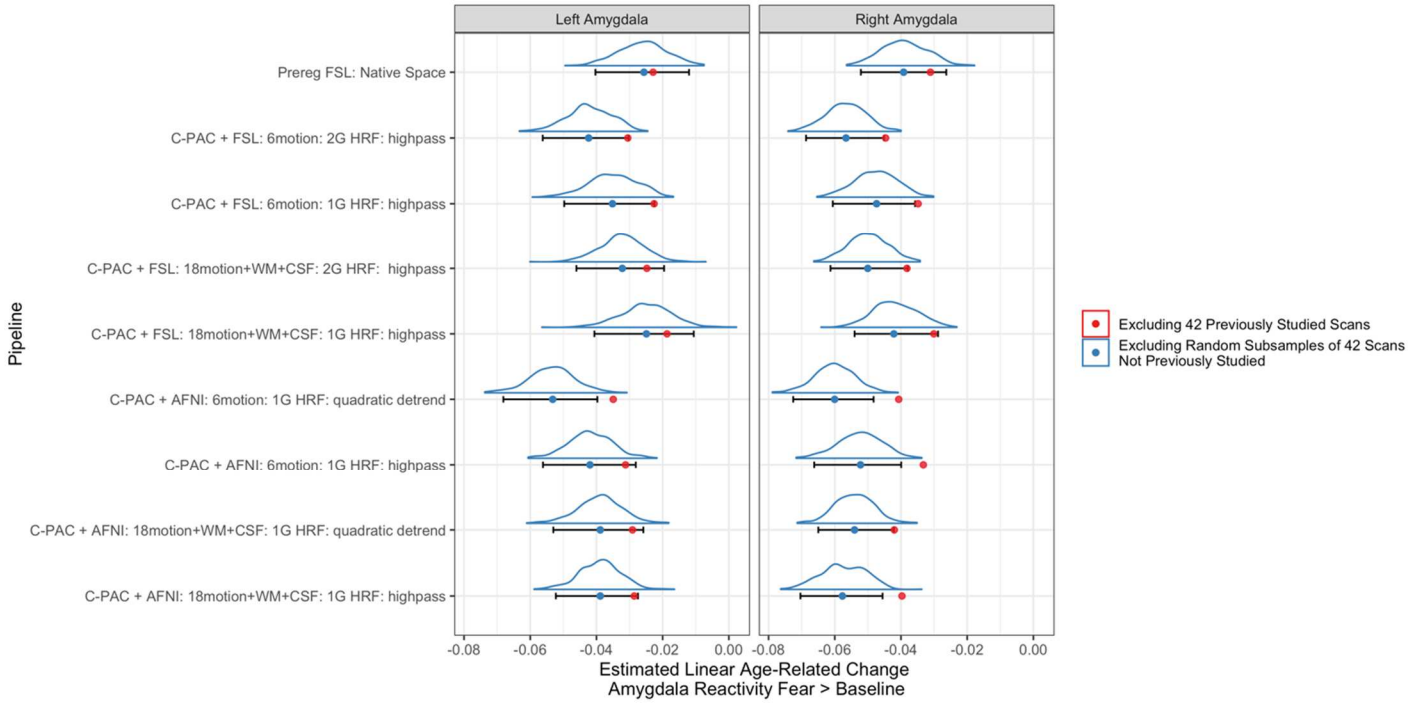


Appendix A Figure 24: Between scan correlations for amygdala reactivity pipelines for fear > neutral.

Dependence of amygdala reactivity age-related change findings on previous work

When we included only scans not previously used as a 'discovery set' to identify voxels changing with age in their reactivity to fear faces (a cluster in the right amygdala; Gee et al., 2013), estimates of age-related change were weaker on average, and the majority of posterior intervals for these estimates included 0. Permutation testing against equally sized samples, where 42 scans not previously studied were excluded from analysis at random (to form a 'null' distribution), indicated that for all pipelines tested, excluding the previously studied scans resulted in numerically weaker (less negative) age-related change (Appendix A Figure 25).

Specifically, for the right amygdala (where age-related change was found in exploratory whole-brain analyses by Gee et al., 2013), age-related change when excluding these previously studied scans was less strong than the vast majority of permutation iterations excluding other scans at random for most pipelines. This indicates that analyses within the present study including these scans may somewhat overestimate age-related change due to partial dependence on the previous selection of the right amygdala by Gee et al. (2013). However, the magnitude of differences in findings between pipelines including versus excluding these previously-studied scans was small, and the vast majority of age-related change estimates are of the same sign regardless of exclusion of these scans. In addition, participants were younger on average by 1.5 years (95% PI [0.02, 3.00]) when studied at timepoint 1 by Gee et al. (2013) compared to other scans analyzed here due to the longitudinal study design. Thus, bias introduced by partially circular analyses here may not substantially alter conclusions across all specifications. Further, we also note that even analyses excluding these 42 participants cannot be considered entirely independent of the previous work, as the present work still examines follow-up scans from many of the same participants and uses the same stimuli.



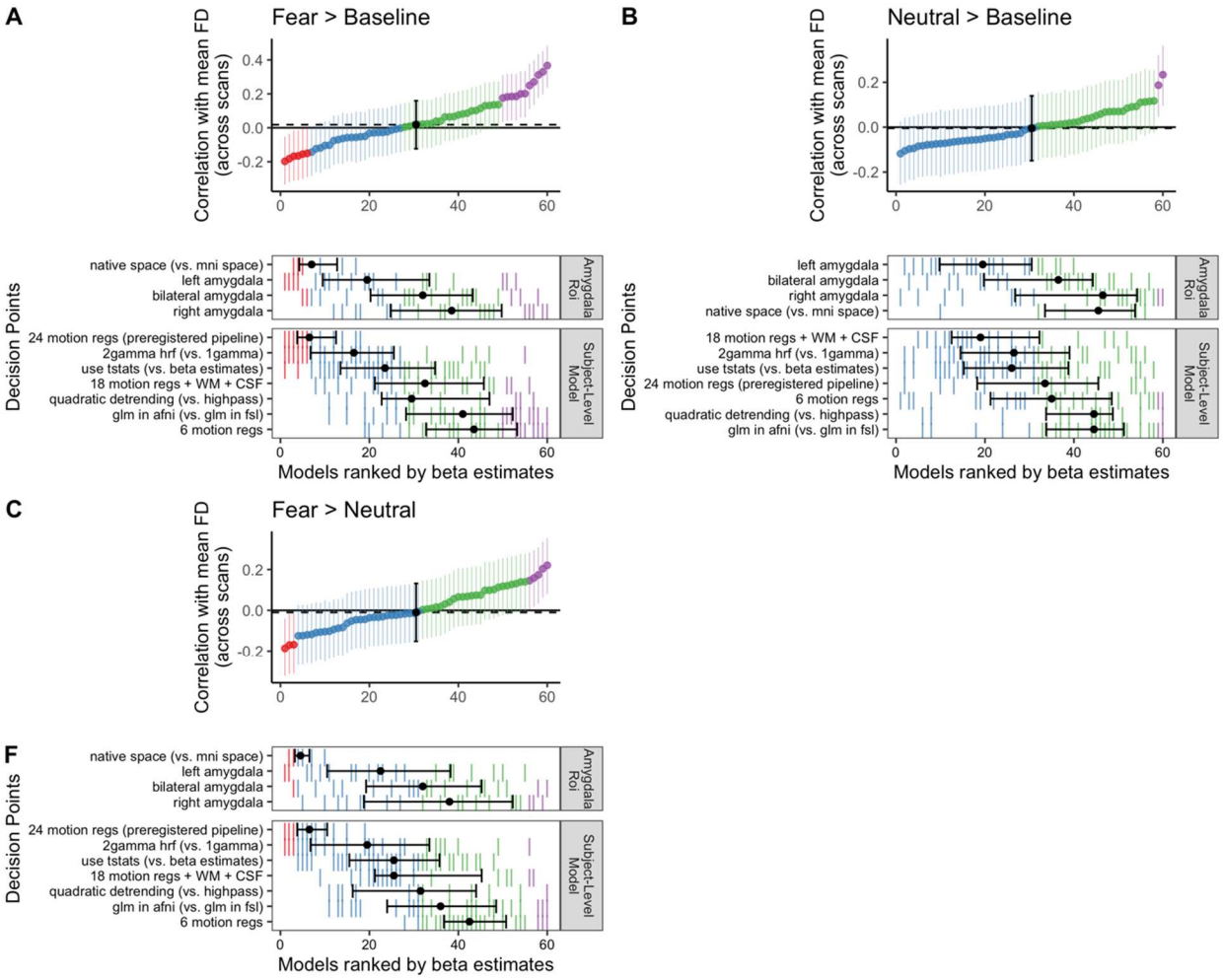
Appendix A Figure 25: Permutation tests for age-related change excluding previously studied scans

Blue distributions are 'null' distributions for age-related change for models of datasets excluding 42 randomly selected scans that had not been previously studied. Error bars indicate 95% confidence intervals based on these distributions, and blue points are the median value. Red points indicate estimate age-related change when excluding the 42 previously studied scans. Red points are always more positive than the blue points (especially for the right amygdala), indicating stronger median negative age-related change when these 42 scans are included than if excluded.

Head Motion & Amygdala Reactivity

We computed product-moment correlations between in-scanner head motion (mean FD) and amygdala reactivity for specifications across preprocessing pipelines and contrasts (see

Appendix A Figure 26). Overall, few specifications resulted in amygdala reactivity estimates that were strongly correlated with head motion, although some estimates were significantly associated with motion (95% confidence interval excluding 0 for each contrast).

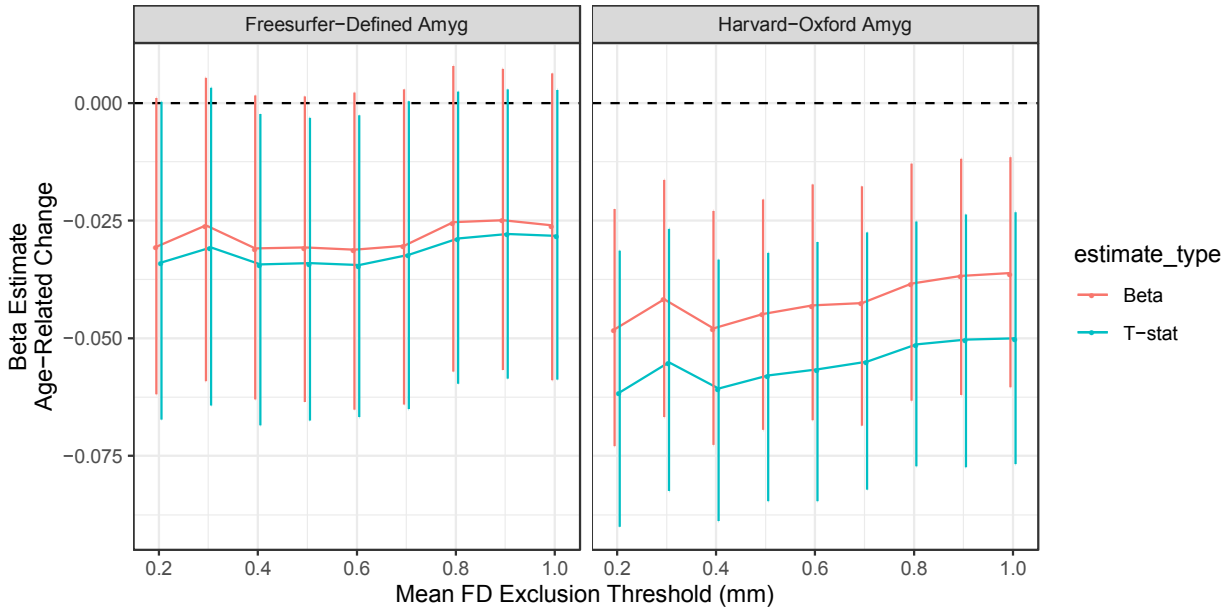


Appendix A Figure 26: Correlations between head motion and amygdala reactivity for each contrast

Plots show specification curves of correlations ranked by their value for the fear > baseline (A), neutral > baseline (B), and fear > neutral (C) contrasts. Color indicates sign of correlation estimates and whether respective 95% confidence intervals include 0 (red = negative excluding

0; blue = negative including 0, green = positive including 0, purple = positive excluding 0, black = median across all specifications).

Our preregistered exclusion criteria (<= 40 TRs with FD >= .9mm) based on in-scanner head motion was relatively lenient, especially compared to recent recommendations for resting-state fMRI preprocessing (Power et al., 2014). To examine whether results were driven by the inclusion of high-motion scans, we systematically varied an inclusion threshold for analysis from mean framewise displacement during the scan of 0.2-1.0mm in increments of 0.1mm. For each dataset based on the different inclusion thresholds, we modeled age-related change (longitudinal model #1) in bilateral amygdala (both Freesurfer-defined in native space and MNI space) reactivity using both t-statistics and beta estimates from the preregistered preprocessing pipeline. We did not observe meaningful differences in estimated age-related change as a function of head motion exclusion thresholds (see Appendix A Figure 27), indicating that age-related change findings in amygdala reactivity reported here are not likely driven purely by high-motion scans.



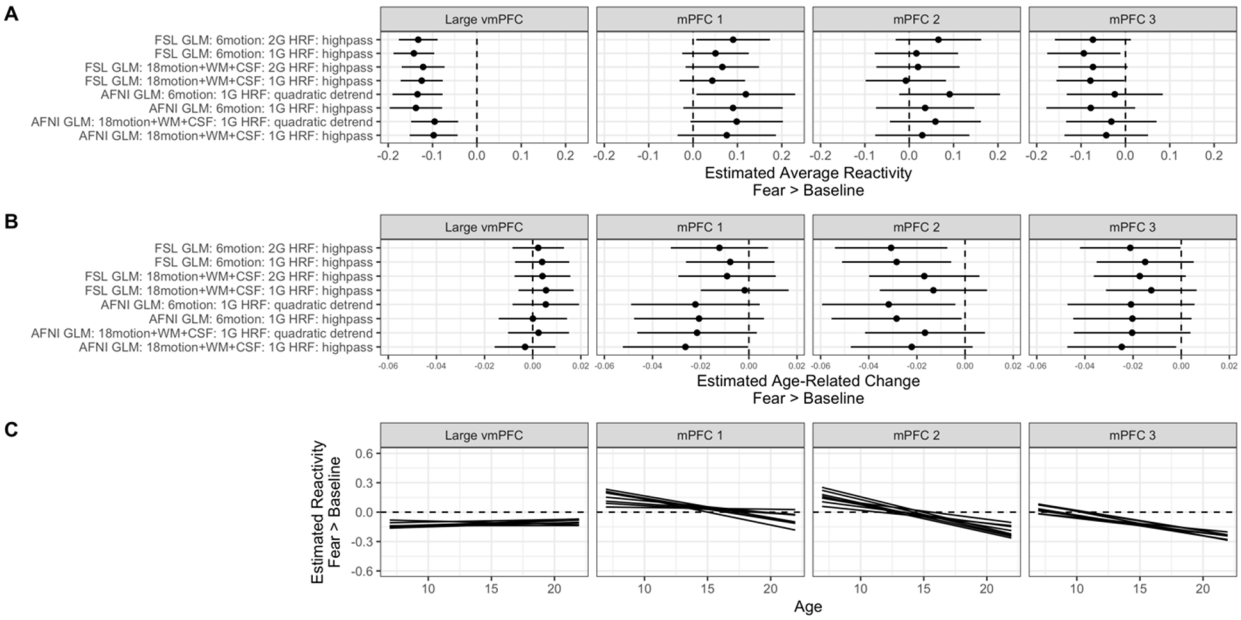
Appendix A Figure 27: Estimated age-related change as a function of mean FD exclusion threshold

The mean FD threshold for exclusion of a scan is shown on the x-axis, and the point estimate and 95% posterior intervals for corresponding age-related change estimates are on the y-axis.

mPFC reactivity: supplemental results

In addition to the amygdala, we also inspected reactivity in each of the 4 mPFC regions for the fear > baseline contrast with separate specification curves. To conserve computational resources, we fit these models using maximum likelihood with the lme4 R package (Bates et al., 2019), rather than fully Bayesian inference. Approximate 95% confidence intervals were constructed from these models by computing the interval ± 2 standard errors from the maximum likelihood estimate. Average reactivity differed by region, such that reactivity to fearful faces in the large vmPFC region was negative (i.e. signal lower than baseline), somewhat negative for mPFC

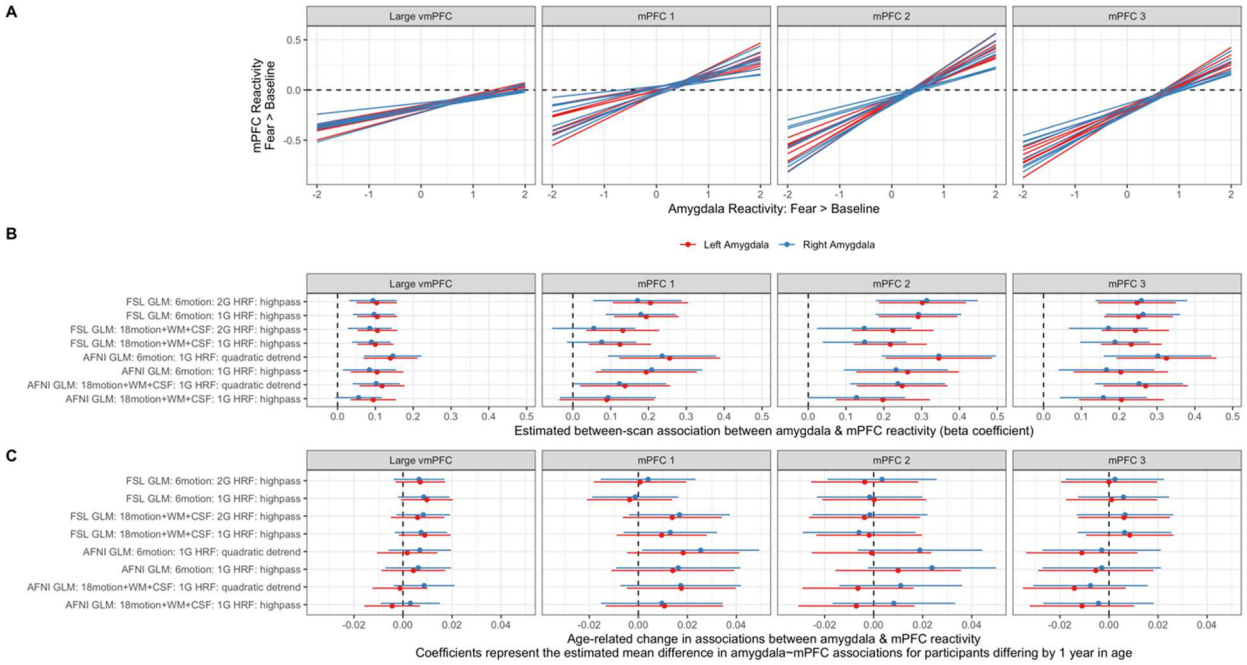
region 3, and somewhat positive for mPFC regions 1-2 (Appendix A Figure 28A). Only in the large vmPFC region was average reactivity for fear > baseline reliably distinguishable from 0 across specifications. While age-related change estimates were rarely distinguishable from 0 for any region, estimates were all negative in sign for mPFC regions 1-3 (Appendix A Figure 28B-C).



Appendix A Figure 28: Group average mPFC reactivity and age-related change for fear faces > baseline

(A) estimated group mean mPFC reactivity for each preprocessing pipeline and ROI. (B) estimated age-related change in mPFC reactivity for each preprocessing pipeline and ROI. For A-B, error bars are approximate 95% confidence intervals. (C) model predictions for estimated fear > baseline mPFC reactivity as a function of age (in years), with different preprocessing pipelines plotted as individual spaghettis for each ROI.

Because Gee et al. (2013) reported that between-participant associations between amygdala and mPFC connectivity were positive among younger children and negative among older youth in a sample from the first timepoint studied here, we examined associations between amygdala and mPFC reactivity similarly in the current longitudinal sample. For these analyses, we also used multilevel linear regression models with the lme4 R package (without random slopes, but with random intercepts and a covariate for head motion). Across preprocessing pipelines, both the left and right amygdala, and all four mPFC regions, fear > baseline amygdala reactivity was positively associated with fear > baseline mPFC reactivity (see Appendix A Figure 29A-B). However, we did not find consistent evidence for age-related differences in associations between amygdala and mPFC reactivity for any mPFC region (Appendix A Figure 29C).



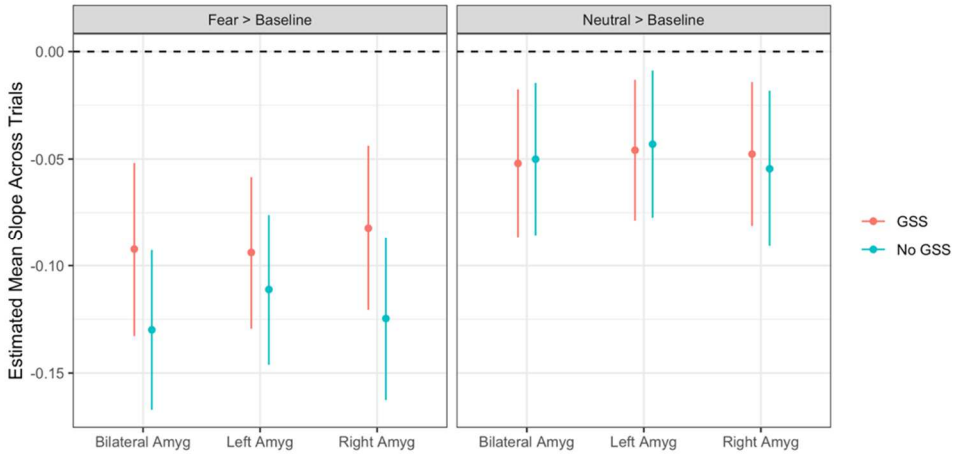
Appendix A Figure 29: Associations between amygdala & mPFC reactivity

(A) Model predictions for fear > baseline mPFC reactivity (y-axis) as a function of fear > baseline amygdala reactivity (x-axis). Individual spaghettis represent models for different preprocessing pipelines for both the left (red) and right (blue) amygdala. (B) Beta coefficients and 95% confidence intervals for associations between amygdala and mPFC reactivity. (C) Beta coefficients and 95% confidence intervals for age-related change in associations between amygdala and mPFC reactivity. Positive terms would represent stronger (more positive) amygdala—mPFC reactivity associations with increasing age.

Within-scan changes in amygdala reactivity: supplemental results

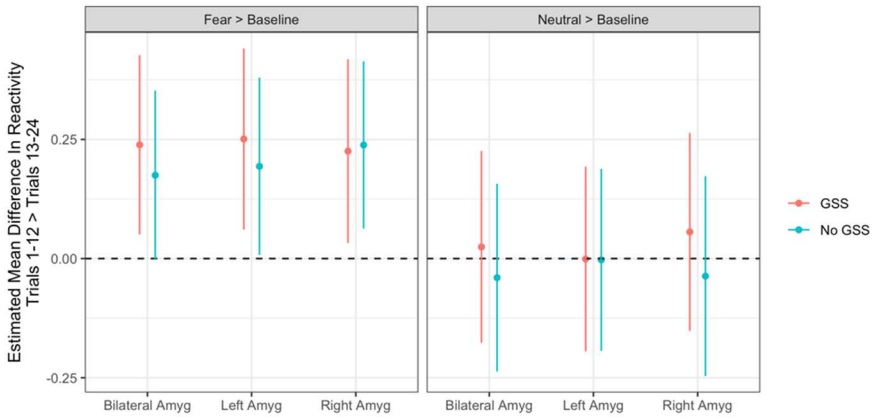
Group average within-scan changes in amygdala reactivity

As with amygdala and mPFC reactivity, we modeled the group average within-scan change in amygdala reactivity using lme4. We estimated both the average slope of amygdala reactivity across trials (such that negative slope indicates linear decreases in amygdala reactivity across trials), and the mean difference between reactivity in trials 1-12 > 13-24 (first half > second half). We computed these group average estimates across bilateral, right, and left amygdala regions, with and without global signal subtraction, and for both the fear > baseline and neutral > baseline contrasts. Average slopes across trials were negative for both fear and neutral faces, although slopes were on average steeper for fear faces (Appendix A Figure 30). While on average, amygdala reactivity was higher for the first half of trials for fear faces across specifications, there were no consistent average differences between trial halves for amygdala reactivity to neutral faces (Appendix A Figure 31).



Appendix A Figure 30: Group average slopes in amygdala reactivity across trials

Negative slopes indicate linear decreases in amygdala reactivity across trials on average. Points display maximum likelihood estimates, and error bars are 95% confidence intervals.



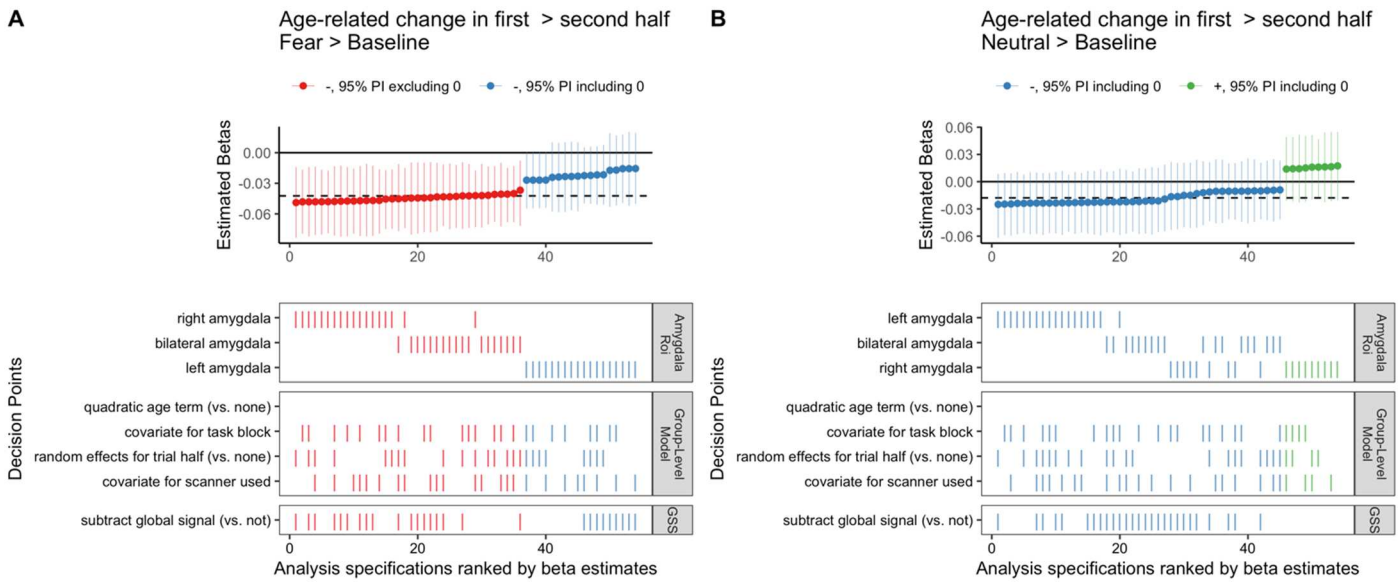
Appendix A Figure 31: Group average differences in amygdala reactivity across first > second half of trials

Positive values indicate higher average amygdala reactivity in the first half of trials. Points display maximum likelihood estimates, and error bars are 95% confidence intervals.

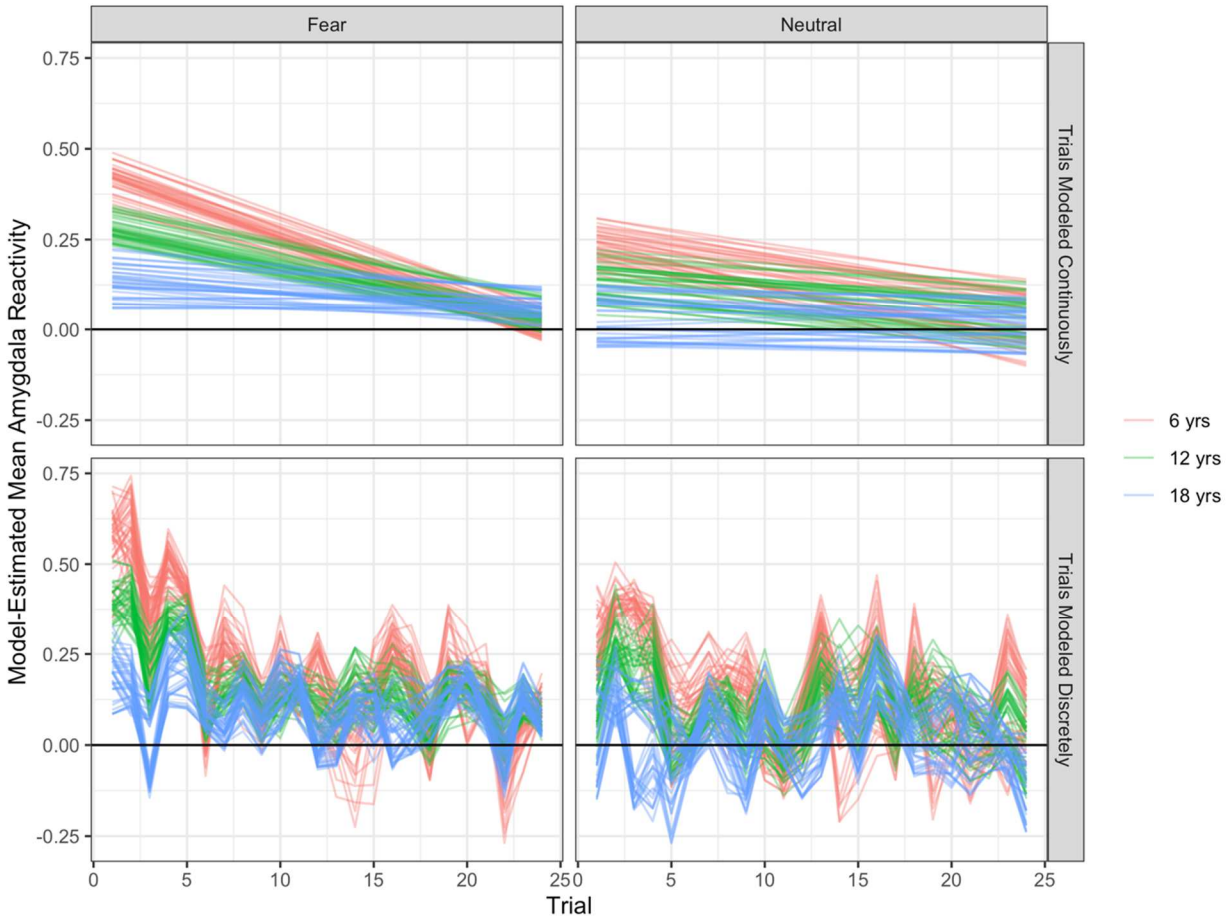
Age-related differences in within-scan amygdala reactivity change

In addition to specification curves for amygdala reactivity slopes across trials (main manuscript Figure 3), we constructed parallel specification curves for age-related change in the difference in amygdala reactivity between trial halves. For the fear > baseline contrast, most specifications found evidence for an interaction between trial half and age (100% in the same direction, 66.7% of posterior intervals excluded 0), such that differences between amygdala reactivity in the first half > second half of trials were more negative (i.e. smaller positive differences, see Appendix A Figure 31) at older ages on average (Appendix A Figure 32A). Posterior intervals for this estimated interaction never excluded 0 for the left amygdala, but always did for the right or bilateral amygdala. For the neutral > baseline contrast, while the majority of specifications (83.3%) found numerically negative change, none of the 95% posterior intervals excluded 0 (Appendix A Figure 32B).

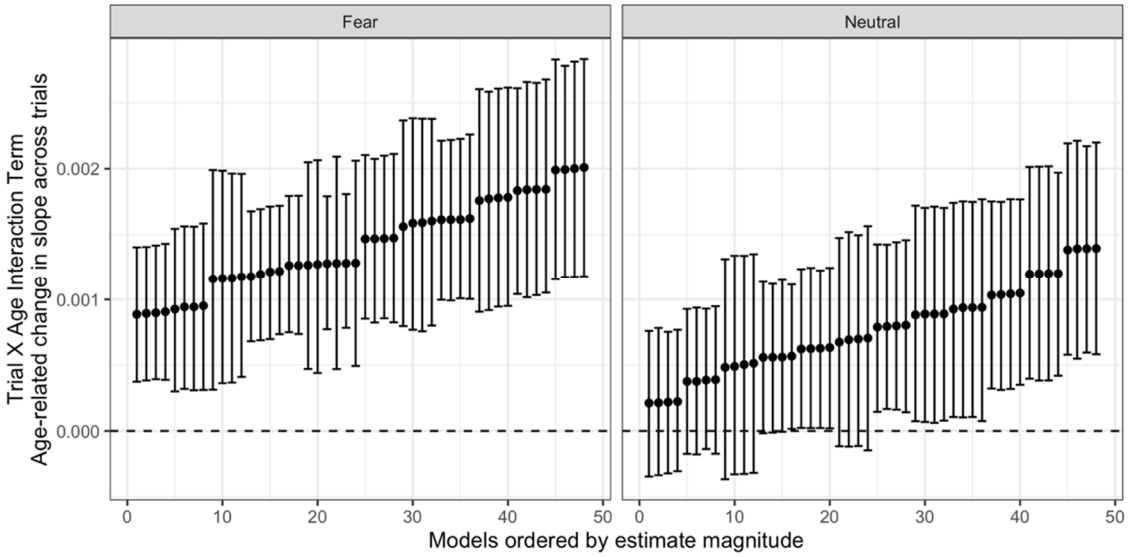
Single-trial models also indicated that for both fear and neutral faces, amygdala responses were larger for early trials for younger children, and more similar across age (though still positive) in later trials (see main manuscript Figure 3C, Appendix A Figure 33). Specification curves for single-trial models indicated that this pattern was somewhat stronger and more robust to analysis choices for fear faces (100% of 95% posterior intervals excluding 0), than for neutral faces (60% of posterior intervals excluding 0, see Appendix A Figure 34). Thus, analyses of slopes across trials, differences between trial halves, and single trial analyses indicated more consistent evidence of age-related change for fear faces than for neutral.



Appendix A Figure 32: Spec. curve for differences across task half in amygdala reactivity age-related change. Specification curves showing parameter estimates and 95% posterior intervals for the estimated interaction term between age*trial half for amygdala reactivity, for the fear > baseline (A) and neutral > baseline (B) contrasts. Negative terms indicate that age-related change is more negative (stronger) during the first 12 trials compared to the last 12 (see main manuscript Figure 3B).



Appendix A Figure 33: Multiverse single-trial model predictions as a function of trial and age Predictions for each single-trial model are plotted as individual 'spaghettis' for both fear trials (left) and neutral trials (right). For illustrative purposes, we plot predictions for an average person at age 6, 12, and 18 years of age. In the top panel, models include terms for linear trial associations, while in the bottom panel, trials are modeled discretely.



Appendix A Figure 34: Spec. curve of age*trial amygdala reactivity interactions from single-trial models

Positive trial*age interaction terms indicate that slopes for within-scan linear changes in amygdala reactivity (as modeled through a single-trial multilevel model) were more positive (i.e. less negative, because slopes were negative on average) at older ages.

gPPI Functional Connectivity Results

Impacts of a deconvolution step on gPPI regressors and estimates

gPPI regressors are interactions formed from the multiplication of the seed timeseries with the stimulus (task) regressors. In order for gPPI to measure 'task-dependent' connectivity, there must be adequate time when stimuli are 'on' versus 'off' such that such an interaction regressor can represent the difference in connectivity between two regions when the stimulus is present versus absent (McLaren et al., 2012). Unfortunately, within rapid event-related design,

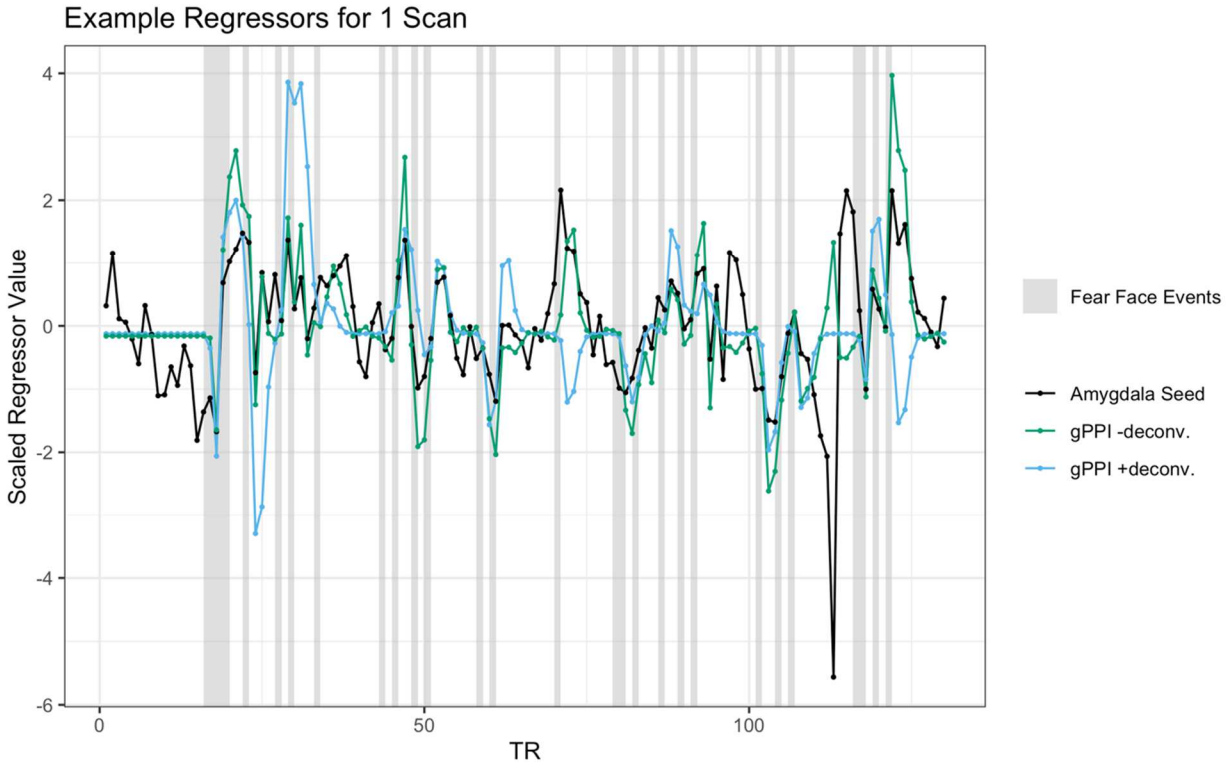
there were few periods of baseline 350ms presentations of either fear or neutral faces (ITI was jittered between 3-9s), other than the initial 20-second fixation period at the beginning of the scan. If it is then unclear which TRs represent connectivity when the stimuli are 'on' versus 'off', estimation of 'task-dependent' connectivity is difficult. Thus, other than the initial 20-second fixation period in our paradigm, resolution between stimulus presentations is low and there are very few moments in which the gPPI regressors are at baseline (i.e. flat, see Appendix A Figure 35 for an example set of regressors). This may especially be a problem for the gPPI regressor without deconvolution, where the stimulus regressor has already been convolved with the HRF before multiplication with the seed timeseries. Because of the slow temporal dynamics of the HRF, the stimulus regressor rarely returns to baseline between events (this would take 15-20s), and the gPPI regressor is then correlated with the seed timeseries.

Indeed, regressors both with and without deconvolution were collinear with the amygdala seed on average, although multicollinearity was more of a problem without deconvolution (Appendix A Figure 36). Thus, gPPI estimates without deconvolution might especially represent associations between brain regions that include "task-independent" signal in addition to connectivity associated particularly with the face stimuli. Within pipelines including a deconvolution step, however, tweaks to regularization methods (adding a lasso penalty) in the 3dTfitter algorithm or up-sampling of the seed timeseries to 0.1s resolution before deconvolution resulted in substantially differing gPPI regressors. As shown in Appendix A Figure 36 (top panel), gPPI regressors for the same scan using different deconvolution methods were only often only weakly associated ($r < .5$). Such differences among regressors all ostensibly using deconvolution indicates that our solutions for the 'underlying neuronal timecourse' (and the

resulting gPPI estimates) may be unreliable due to their sensitivity to such changes in the deconvolution pipeline.

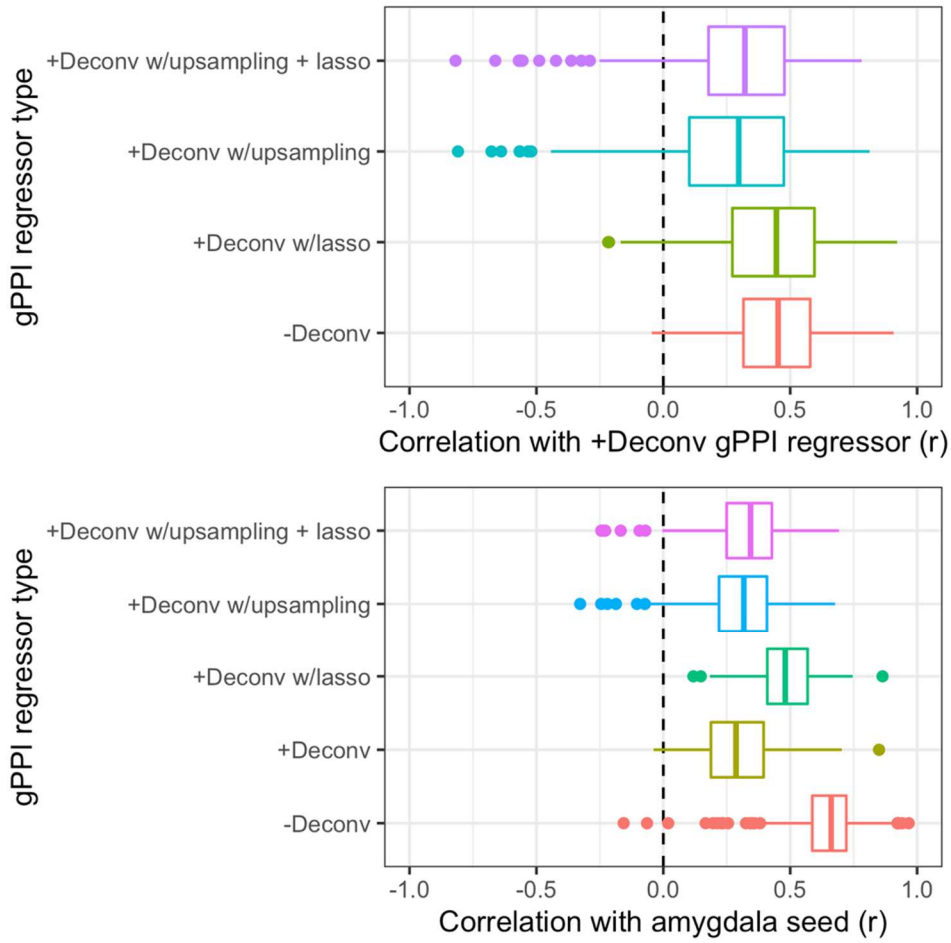
We then asked how similar amygdala gPPI results would be at a voxelwise level when using a pipeline with versus without deconvolution. We compared voxelwise patterns between t-statistic maps for fear > baseline, fear > neutral, and neutral > baseline gPPI with versus without deconvolution for the same scans. Patterns were overall positively correlated across for all mPFC regions (as well as whole-brain patterns), but varied significantly such that the median correlation between patterns with versus without deconvolution was never above r = .5 for the fear > baseline or neutral > baseline contrasts (Appendix A Figure 37). While patterns were slightly more similar for the fear > neutral contrast, that such correlations were only moderate indicated substantial differences in patterns of results across voxels between pipelines with versus without deconvolution. The higher variability in similarity values across mPFC regions 1-3 is likely because these ROIs were much smaller, and thus correlations were calculated across far fewer voxels. These substantial differences when comparing amygdala gPPI results with versus without deconvolution for the same scan likely play a major role in the discrepancies in findings for age-related change in amygdala—mPFC gPPI. Although we did not test other task paradigms here, we speculate that gPPI analyses with block designs or event-related designs with

longer ITIs (20+ seconds) may be more successful at estimating task-dependent connectivity.



Appendix A Figure 35: Example gPPI and amygdala seed regressors for one scan

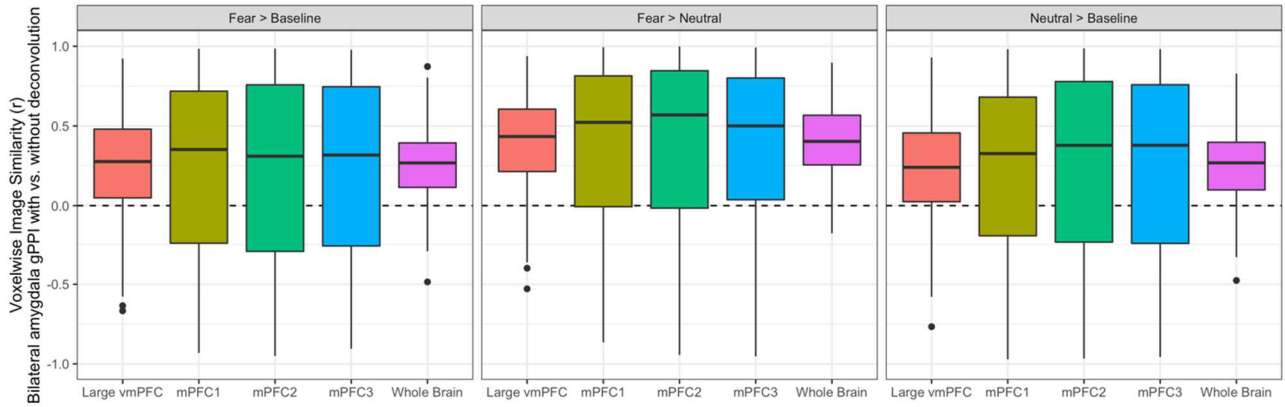
Timecourses for the average timeseries of the amygdala seed (black), gPPI regressor without deconvolution (green), and gPPI regressor with deconvolution (blue) for an example participant. TRs where fear face stimuli are presented are highlighted in grey. gPPI regressors, especially the regressor without deconvolution, are rarely flat other than at the beginning of the task, because there are few temporal gaps between fear face stimuli of more than 10-15s. Thus, both gPPI regressors, but especially the one without deconvolution, tend to be correlated with the seed timecourse.



Appendix A Figure 36: Correlations between gPPI regressors, and between gPPI regressors and the seed

Top: boxplots show distributions of correlations across between different versions of gPPI regressors for the same scan for all scans. +Deconv represents the deconvolved regressor without up-sampling or lasso regularization, which was used for pipelines in the main manuscript. These +Deconv regressors were about equally similar to -Deconv regressors (regressors used in the main manuscript without deconvolution) as they were to +Deconv regressors with lasso regularization. +Deconv regressors were even less similar to +Deconv regressors with up-sampling, or up-sampling and lasso regularization. Bottom: boxplots show distributions of

correlations between different versions of gPPI regressors with the amygdala seed timeseries across all scans.

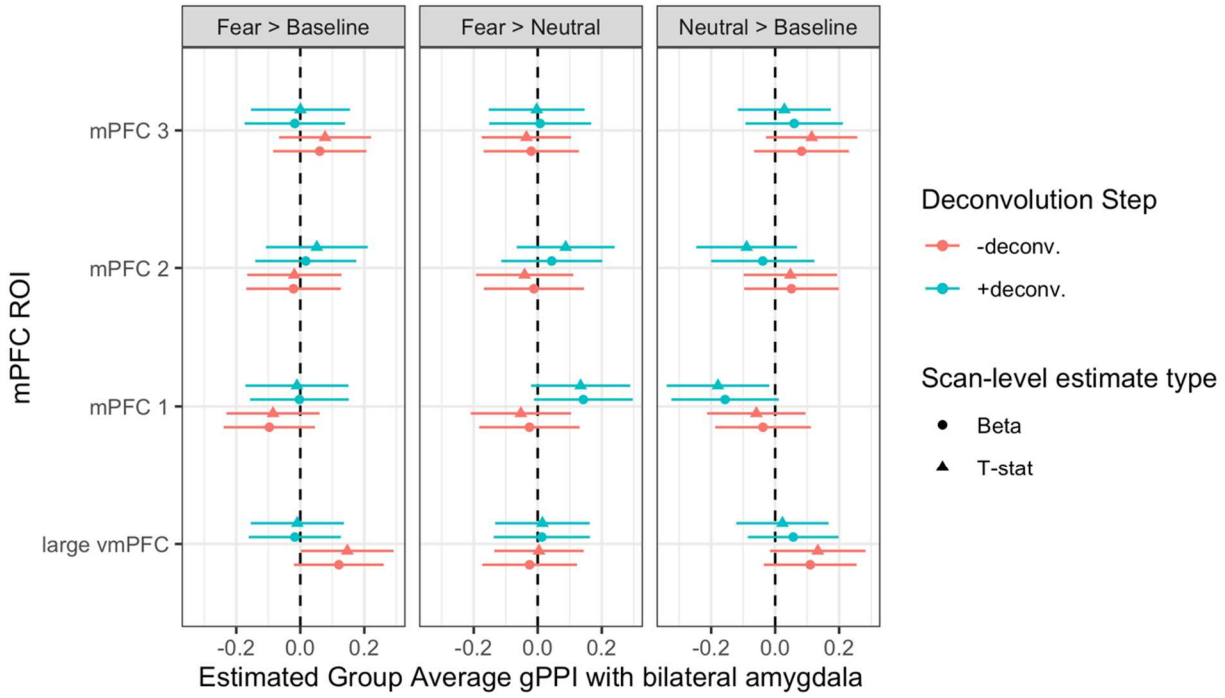


Appendix A Figure 37: Voxelwise image similarities of gPPI estimates with vs. without deconvolution

Values on the y-axis represent product-moment correlations between t-statistic maps for each ROI with versus without deconvolution.

Group mean amydala-mPFC gPPI estimates

To preserve computational resources, all models for mean amygdala—mPFC gPPI were fit using the lme4 R package with intercepts allowed to vary by participant. Overall, we did not see consistent evidence for group mean task-dependent amygdala—mPFC connectivity distinct from 0 for any contrast or ROI (see Appendix A Figure 38). As discussed previously, however, the lack of detected group average task-dependent connectivity may owe more to the fact that the task paradigm was ill-suited for estimation of gPPI than absence of true connectivity.



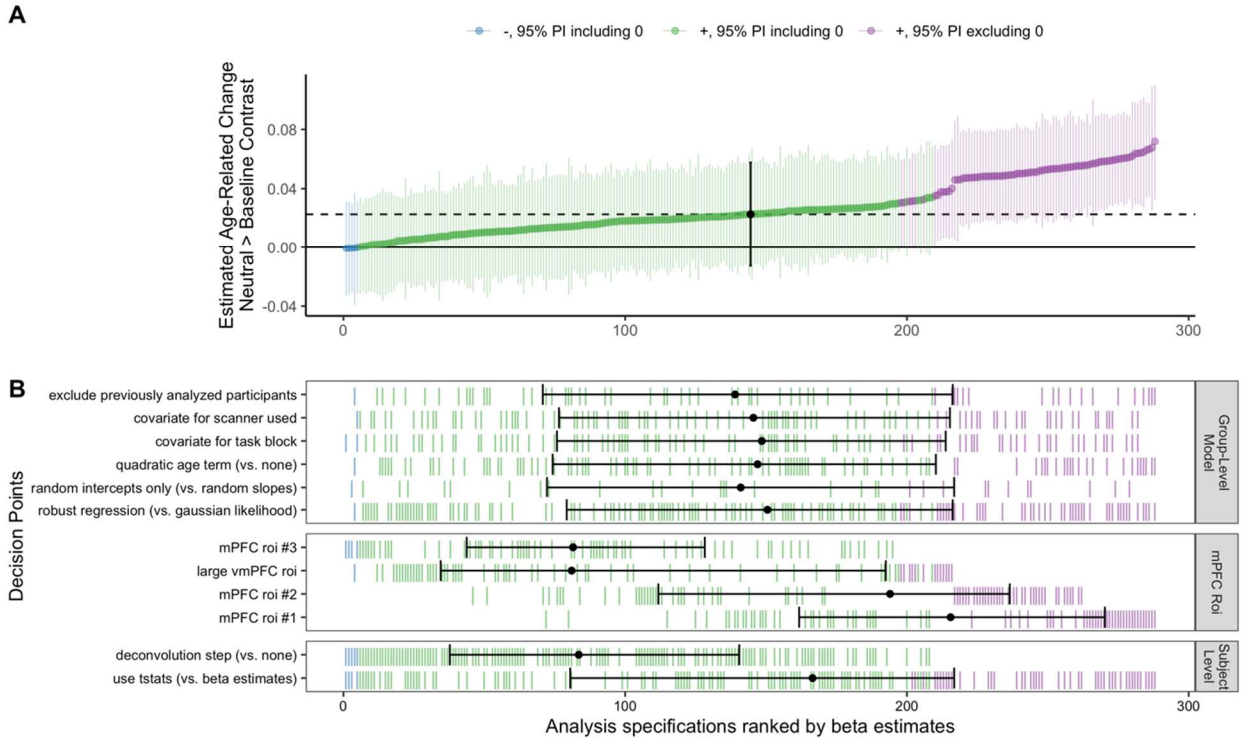
Appendix A Figure 38: Group average amygdala—mPFC gPPI estimates for each contrast and ROI

The x-axis represents estimated average task-dependent amygdala—mPFC connectivity for each ROI (on the y-axis) and contrast (left = fear > baseline, middle = fear > neutral, right = neutral > baseline). Pipelines with deconvolution are represented in blue, and without deconvolution in red.

Multiverse analyses of age-related change in amygdala—mPFC gPPI

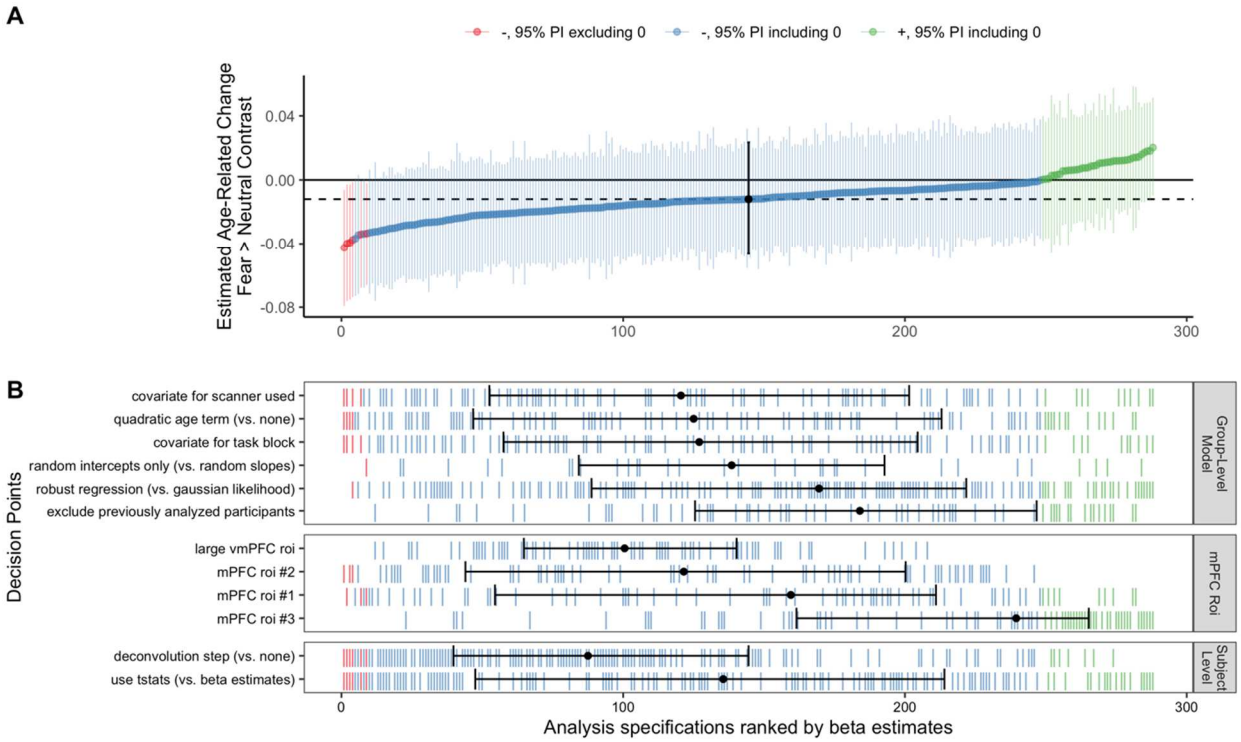
In addition to the constructing specification curves for age-related change in amygdala—mPFC gPPI for the fear > baseline contrast as reported in the main manuscript (see Figure 4), we constructed parallel specification curves for the neutral > baseline and fear > neutral contrasts. For the neutral > baseline contrast the vast majority of pipelines (98.2%) found positive age-related change, though only 29.5% of pipelines estimated such change with a posterior interval

excluding 0 (Appendix A Figure 39). Similar to the fear > baseline contrast, pipelines with a deconvolution step tended to find less positive age-related change. For the fear > neutral contrast, 86.1% of pipelines estimated numerically negative age-related change, though only 2.1% of pipelines did so with a posterior interval excluding 0 (Appendix A Figure 40). Thus, across all contrasts there was no consistent evidence of age-related change in amygdala—mPFC gPPI.



Appendix A Figure 39: Spec. curve for age related change in amygdala—mPFC gPPI for neutral > baseline

A: Points represent estimated linear age-related change in amygdala—mPFC gPPI for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.



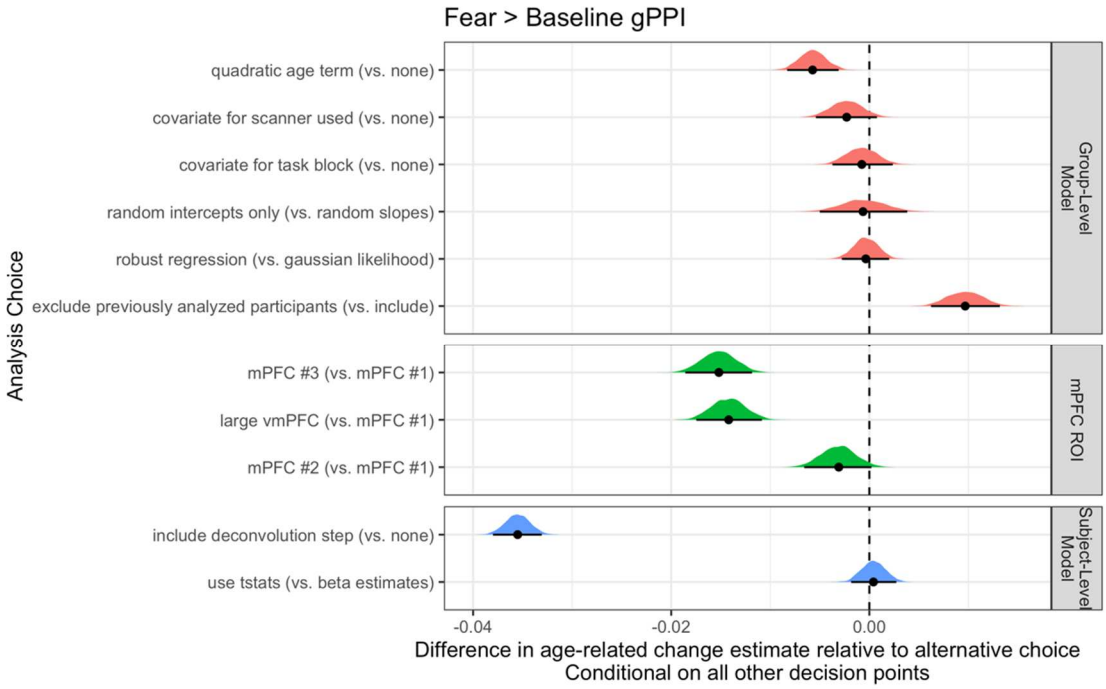
Appendix A Figure 40: Spec. curve for age related change in amygdala—mPFC gPPI for fear > neutral

A: Points represent estimated linear age-related change in amygdala—mPFC gPPI for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.

Impacts of analysis choices on age-related change estimates for amygdala—mPFC gPPI

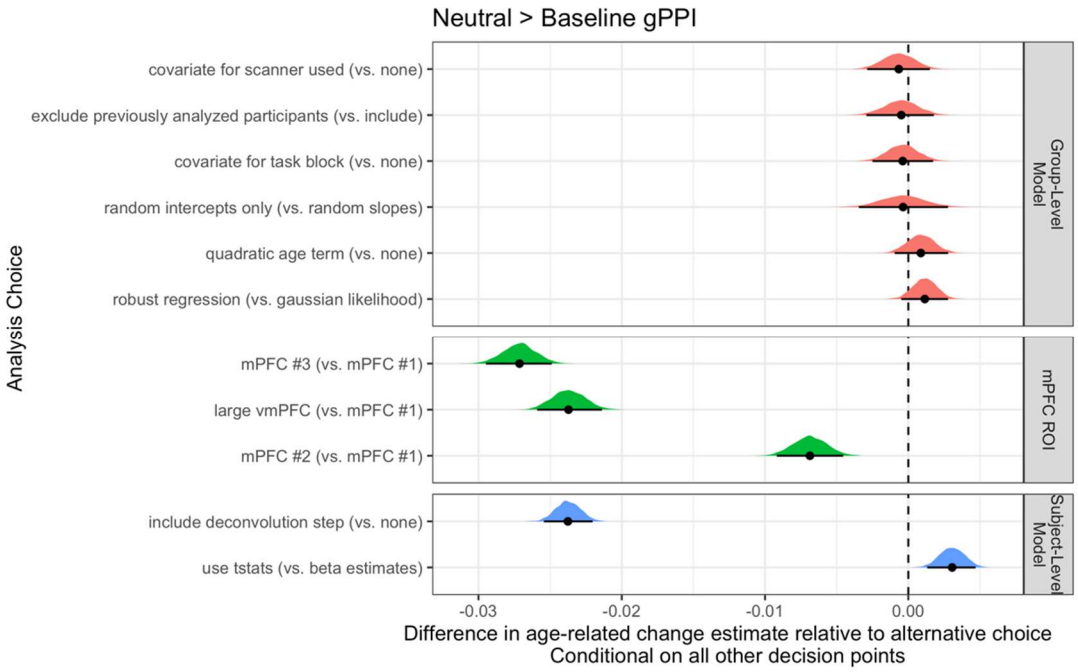
As with amygdala reactivity, we explored the impacts of gPPI analysis choices on estimates of linear age-related change, conditional on all other decision points. For the fear > baseline contrast, whether to use a deconvolution step or not made by far the biggest impact on estimated age-related change (see Appendix A Figure 41). Deconvolution also impacted

estimates for the neutral > baseline contrast (Appendix A Figure 42) and fear > neutral contrast (Appendix A Figure 43). The chosen mPFC ROI also had a large impact on estimated age-related change.



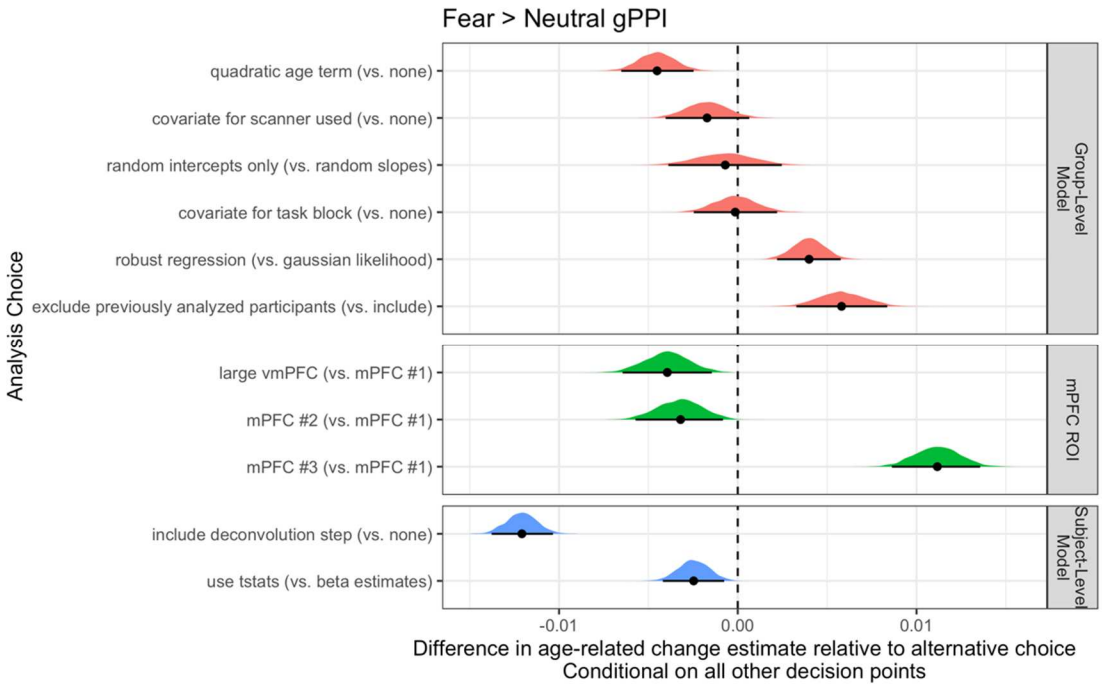
Appendix A Figure 41. Fork impacts on age-related change for fear > baseline amygdala—mPFC gPPI

Posterior distributions and 95% posterior intervals are shown, representing the average difference in linear age-related change estimates relative to the alternative choice.



Appendix A Figure 42: Fork impacts on age-related change for neutral > baseline amygdala—mPFC gPPI

Posterior distributions and 95% posterior intervals are shown, representing the average difference in linear age-related change estimates relative to the alternative choice.



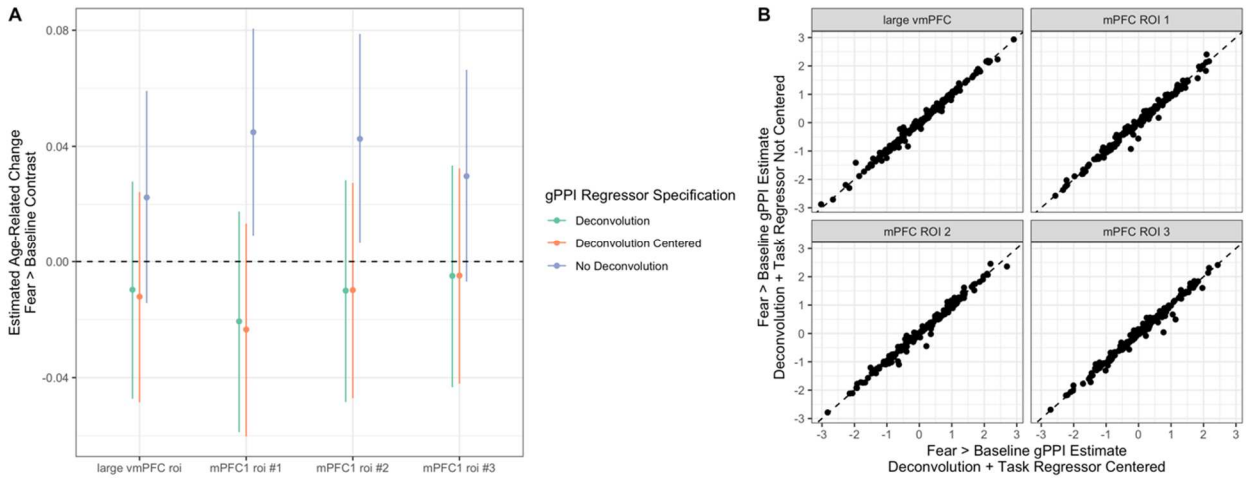
Appendix A Figure 43: Fork impacts on age-related change for fear > neutral amygdala—mPFC gPPI

Posterior distributions and 95% posterior intervals are shown, representing the average difference in linear age-related change estimates relative to the alternative choice.

Impacts of regressor centering on age-related change in amygdala—mPFC gPPI

As previous work has recommended centering the task regressor in gPPI models using deconvolution (Di et al., 2017), we investigated whether age-related change gPPI results with deconvolution differed as a function of this centering. Overall, regressor centering had little impact on linear age-related change estimates, with deconvolution and choice of mPFC ROI having relatively more influence on regression estimates (Appendix A Figure 44A). Further, scan-level estimates for the fear > baseline contrast were highly similar between pipelines where

the task regressor was centered before creating the gPPI regressor and pipelines where no such centering was done (Appendix A Figure 44B). The similarity of scan-level estimates indicated that this centering step had little impact on individual gPPI estimates in this instance.



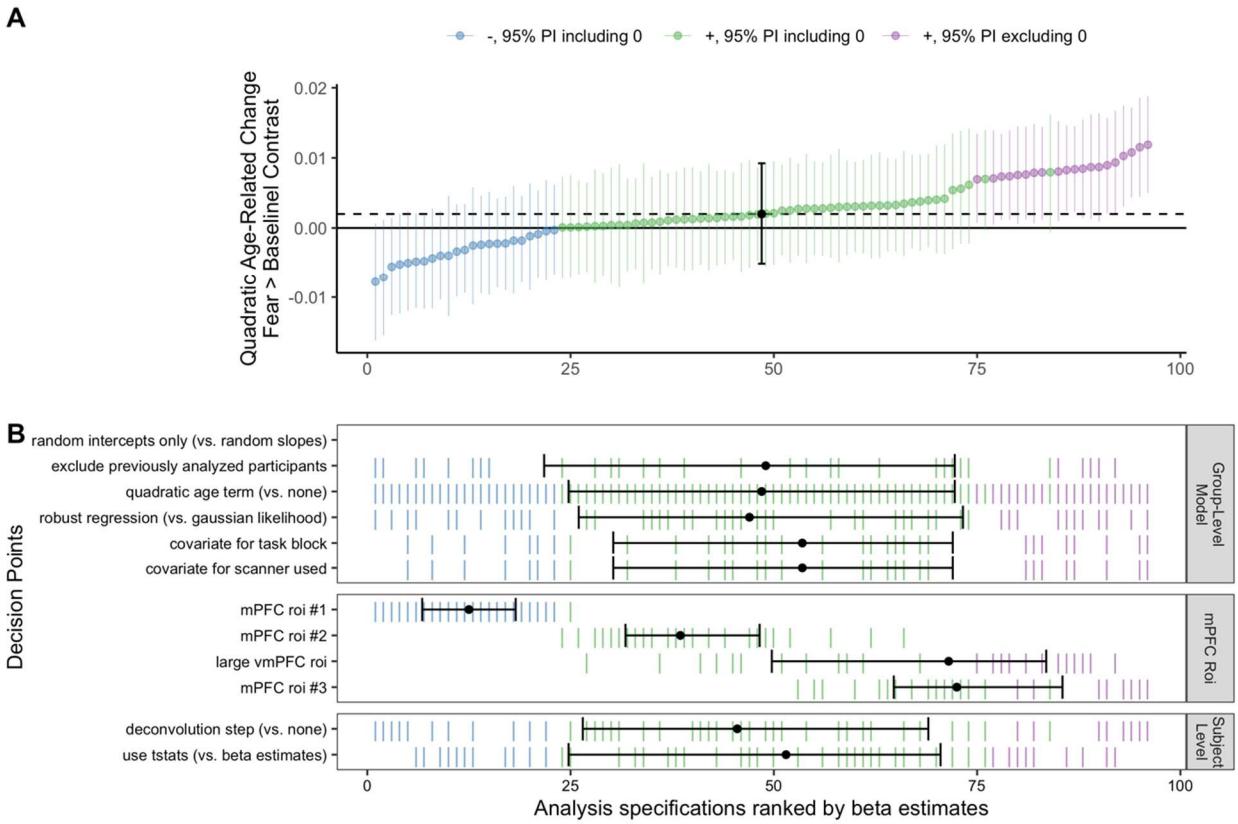
Appendix A Figure 44: Impacts of centering the task regressor in gPPI models with deconvolution

A: Posterior distributions and 95% posterior intervals are shown for age-related change estimates in amygdala—mPFC gPPI functional connectivity as a function of the gPPI regressor specification. Models with deconvolution and without centering of the task regressor (green) demonstrated highly similar estimated age-related change to models with deconvolution and centering the task regressor (orange). Models without deconvolution (purple), by comparison, showed differing estimates of age-related change. Models displayed are for the fear > baseline contrast. B: Direct comparison of scan-level estimates for fear > baseline amygdala—mPFC gPPI with deconvolution and with the task regressor centered (x-axis) versus not centered (y-axis). Panels separate gPPI estimates by mPFC ROI.

Non-linear age-related changes in amygdala—mPFC gPPI

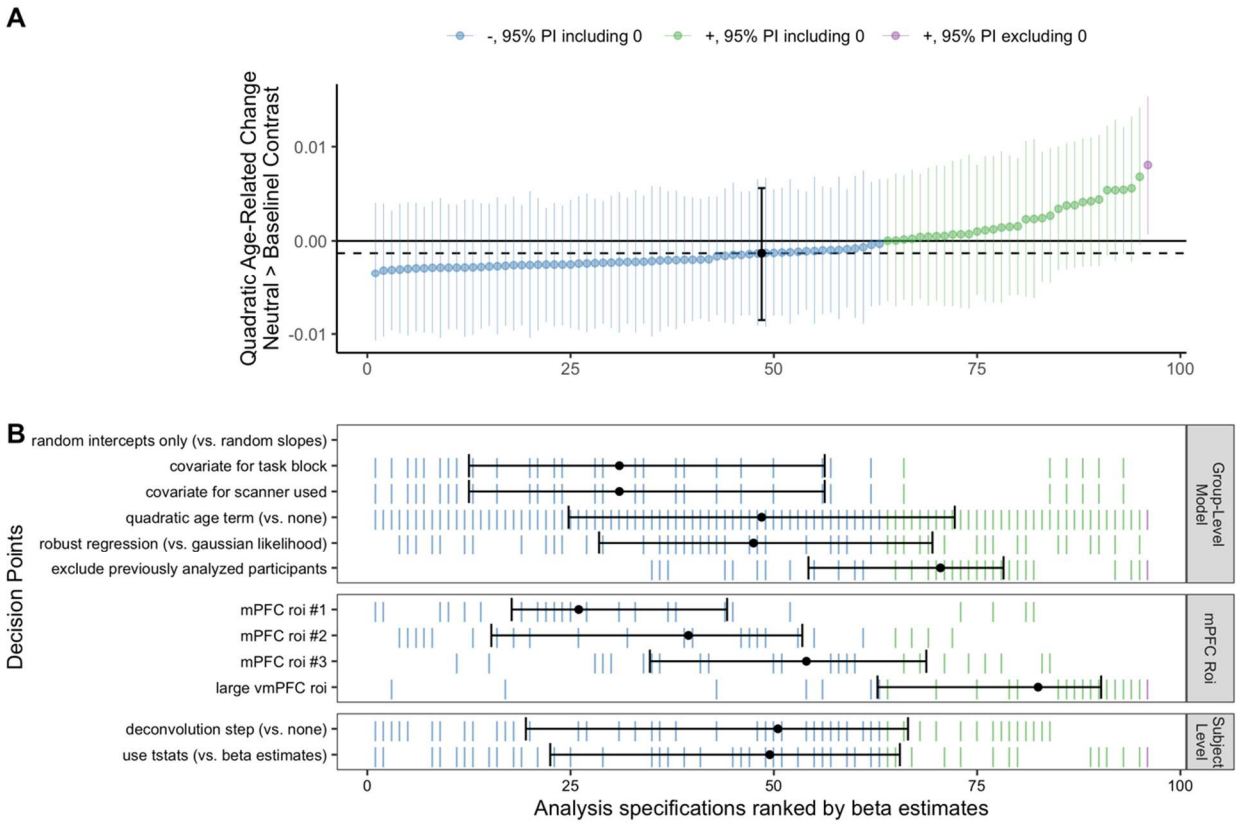
We constructed specification curves for quadratic age-related change parameters across all models including quadratic terms for the fear > baseline (Appendix A Figure 45), neutral > baseline (Appendix A Figure 46), and fear > neutral (Appendix A Figure 47) amygdala—mPFC gPPI. Across all contrasts, few specifications (20.8% for fear > baseline, 1.0% for neutral > baseline, 8.3% for fear > neutral) estimated quadratic terms distinguishable from 0. Quadratic fits also varied considerably in sign for each contrast, such that there was not consensus on 'U-shaped' or 'inverse U-shaped' change. Thus, while the current study may not have been adequately powered to estimate quadratic age-related change, we did not find consistent evidence for either peaks or troughs in amygdala—mPFC gPPI.

We also constructed inverse age models for age-related change in amygdala—mPFC gPPI (Appendix A Figure 48). As with linear models, such models indicated that a deconvolution step largely influenced age-related change estimates for the fear > baseline contrast, often flipping the sign from positive change (without deconvolution) to negative change (with deconvolution, Appendix A Figure 48). However, deconvolution had a relatively smaller impact for the neutral > baseline and fear > neutral contrasts.



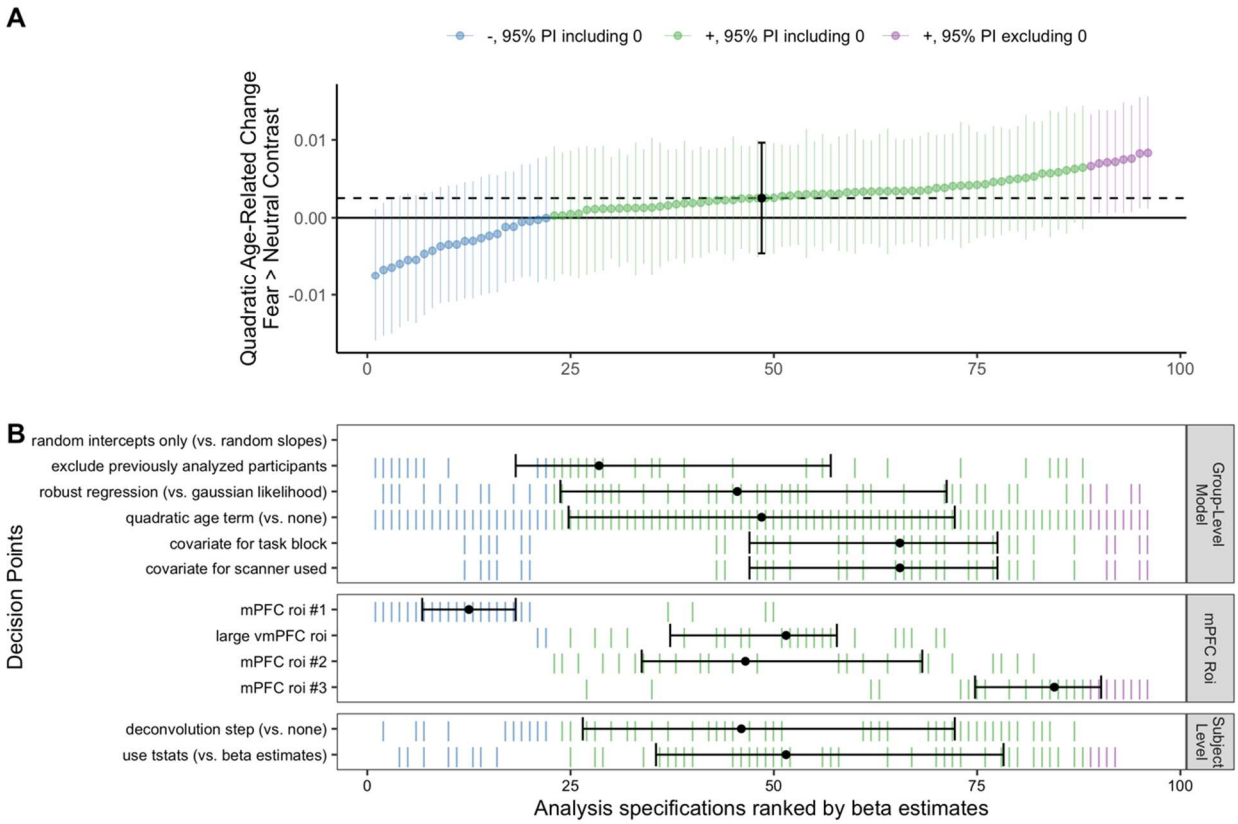
Appendix A Figure 45: Spec. curve for quadratic age-related changes in fear > baseline amygdala—mPFC gPPI

A: Points represent estimated quadratic age-related change in amygdala—mPFC gPPI for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.



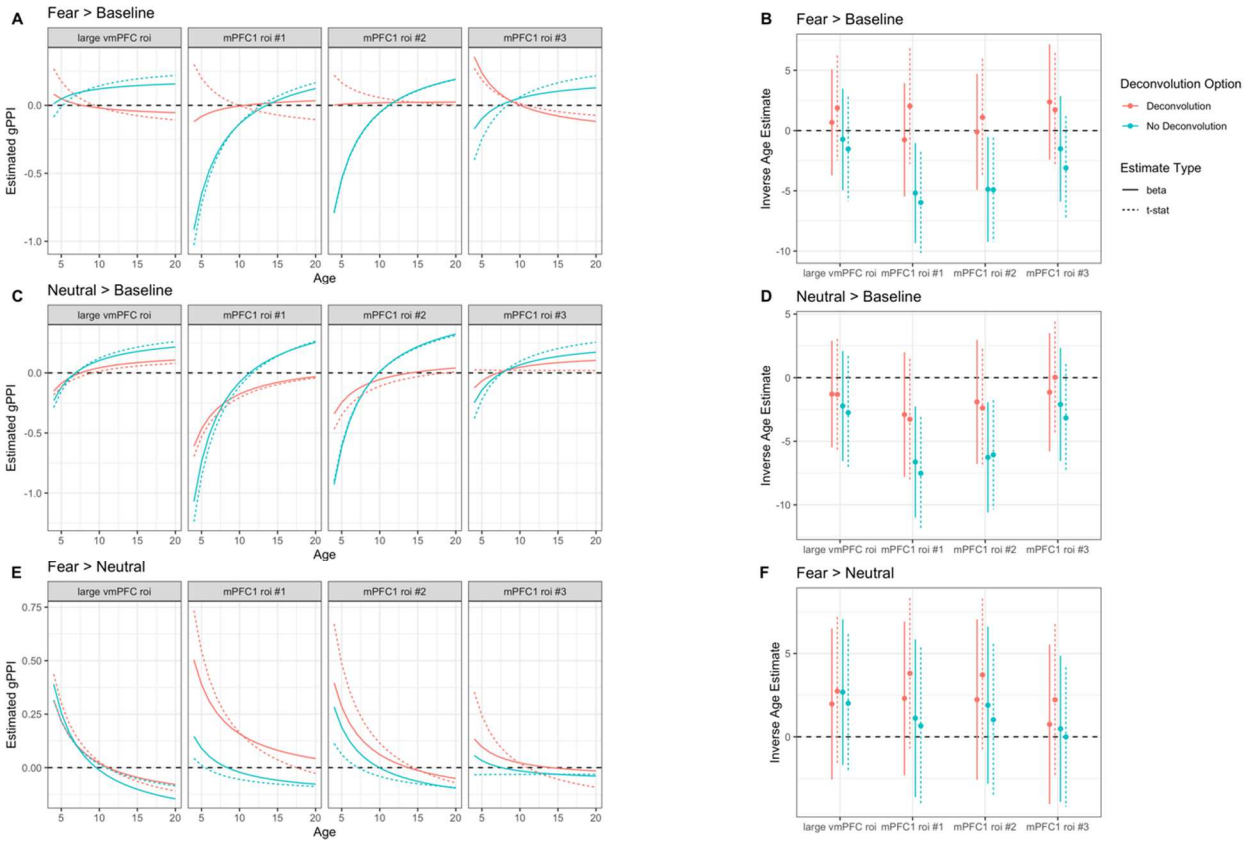
Appendix A Figure 46: Spec. curve for quadratic age-related changes in neutral > baseline amygdala—mPFC gPPI

A: Points represent estimated quadratic age-related change in amygdala—mPFC gPPI for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.



Appendix A Figure 47: Spec. curve for quadratic age-related changes in fear > neutral amygdala—mPFC gPPI

A: Points represent estimated quadratic age-related change in amygdala—mPFC gPPI for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.



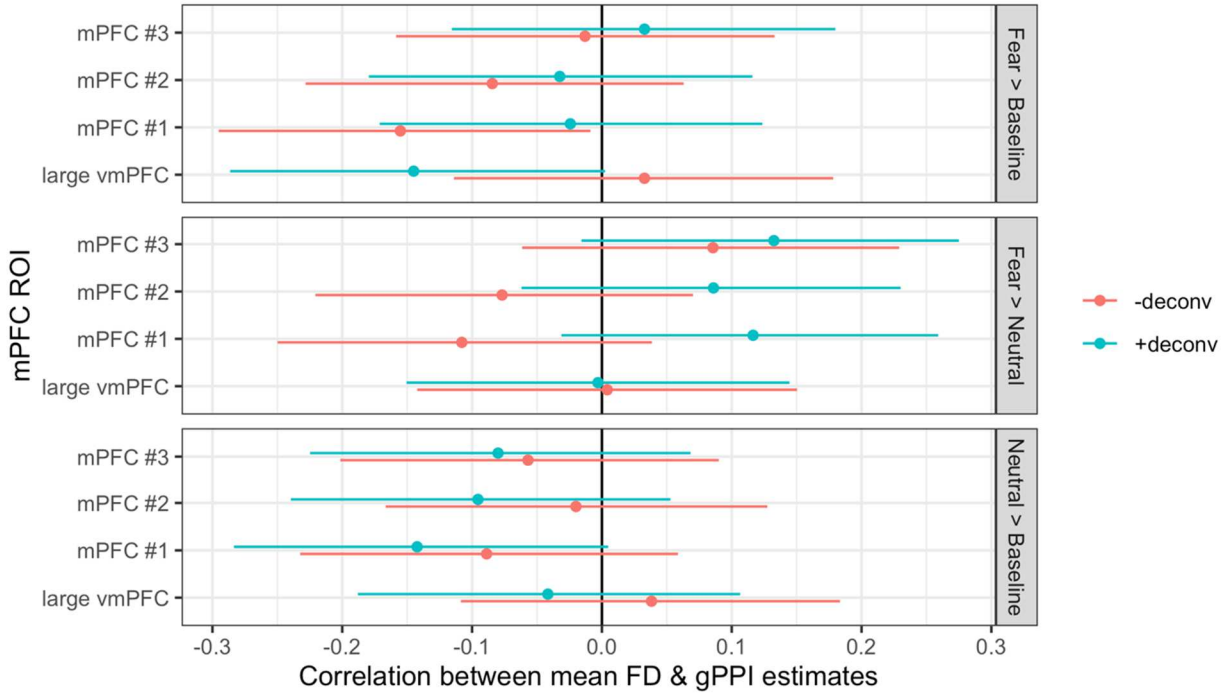
Appendix A Figure 48: Inverse age models for amygdala—mPFC gPPI

Left panels show fitted model predictions for inverse age models for the fear > baseline (A, top), neutral > baseline (C, middle), and fear > neutral (E, bottom) contrasts. Specifications including a deconvolution option are plotted in red, and without deconvolution in blue. Specifications using beta estimates are plotted with filled lines and with t-stats using dotted lines. Right panels show beta estimates for corresponding models for each contrast. Positive estimates for inverse age indicate decreases in amygdala—mPFC gPPI as a function of age, and vice-versa.

Correlations between head motion and amygdala—mPFC gPPI estimates

We computed product-moment correlations between in-scanner head motion (mean FD) and amygdala—mPFC gPPI across pipelines (deconvolution versus none), contrasts, and ROIs (see

Appendix A Figure 49). Overall, head motion was not strongly correlated with gPPI estimates, such that few 95% confidence intervals for correlations excluded 0.



Appendix A Figure 49: Correlations between mean FD & gPPI estimates across scans

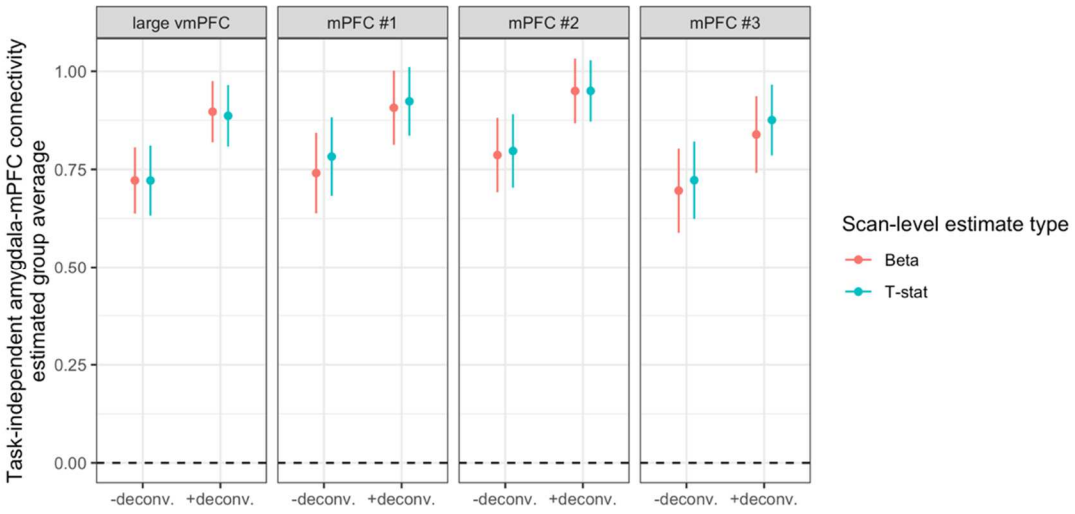
For each contrast, pipeline (+deconv. versus -deconv), and ROI, points show estimated product-moment correlations and error bars represent 95% confidence intervals.

Task-independent amygdala—mPFC connectivity estimates from gPPI models

Within the gPPI model, the association between the seed timeseries (or 'physiological' term) and target voxel has been conceptualized as representing 'task-independent' functional connectivity (Greene et al., 2020). Although we cannot be sure that such measurements are truly 'task-independent' without analysis of tasks beyond the current study, we used these estimates to explore amygda-mPFC functional connectivity while controlling for task-induced variance. For all mPFC ROIs, such task-independent connectivity with the amygdala was positive on average

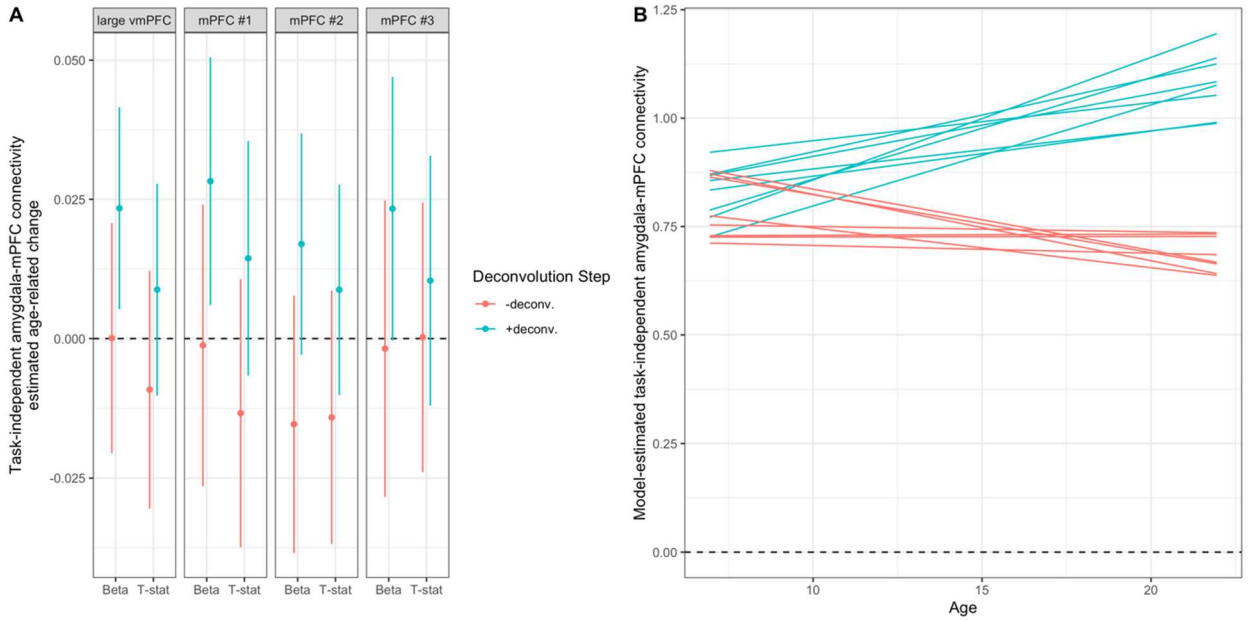
across all participants (Appendix A Figure 50). The positive task-independent amygdala—mPFC connectivity found here may be an overestimate, however, as gPPI pipelines did not include a global signal correction (Power et al., 2017).

In addition, pipelines with deconvolution found age-related increases in task-independent amygdala—mPFC connectivity, while pipelines without deconvolution found age-related decreases (Appendix A Figure 51B). Few 95% confidence intervals excluded 0 for age-related change for either set of pipelines however (Appendix A Figure 51A).



Appendix A Figure 50: Group average task-independent amygdala—mPFC connectivity

Points show estimates and error bars represent 95% confidence intervals for each pipeline and ROI.



Appendix A Figure 51: Age-related change in task-independent amygdala—mPFC connectivity

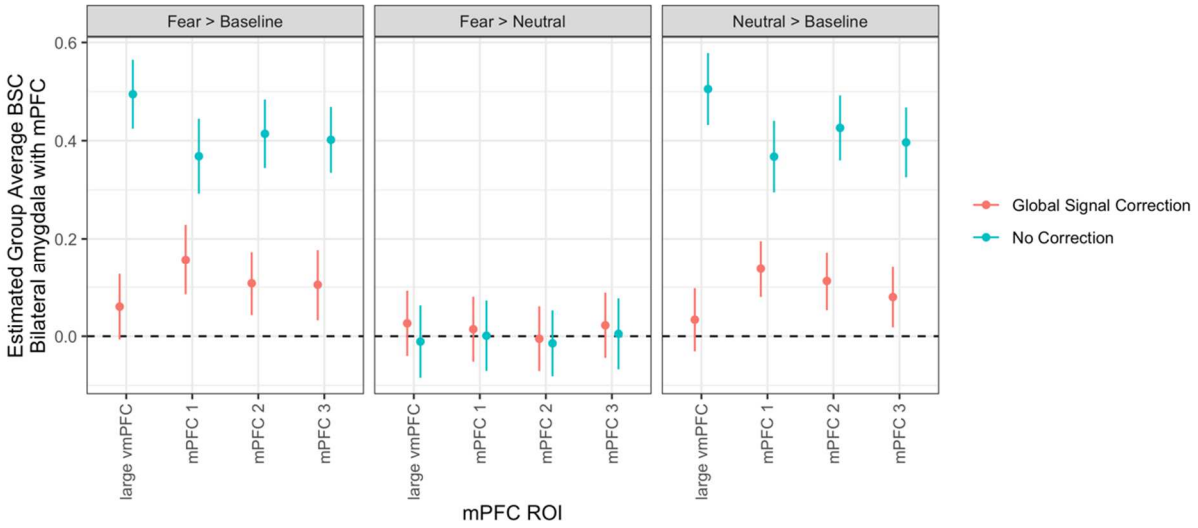
Age-related change coefficients in task-independent amygdala—mPFC connectivity across methods and ROIs (A). The y axis indicates estimated linear age-related change in amygdala—mPFC connectivity. In (B), model predictions as a function of age are plotted as spaghettis in red for pipelines without deconvolution, and blue for pipelines with deconvolution.

BSC Functional Connectivity Results

Group mean amygdala—mPFC BSC

To preserve computational resources, all models for mean amygdala—mPFC BSC were fit using the lme4 R package with intercepts allowed to vary by participant. For both the fear > baseline and neutral > baseline contrasts, we found positive amygdala—mPFC connectivity for all pipelines without a global signal correction, and positive (yet weaker) connectivity for all

pipelines with a correction (Appendix A Figure 52). We did not find average differences between amygdala—mPFC BSC for the fear > neutral contrast (Appendix A Figure 52).



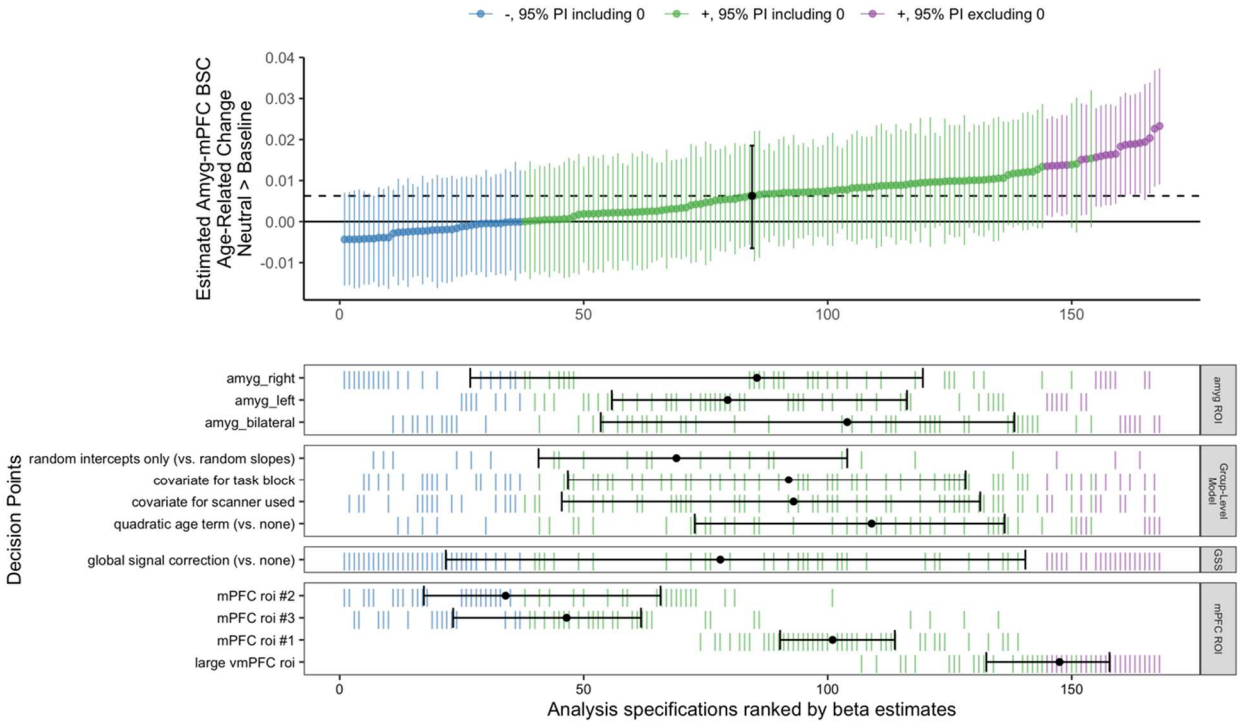
Appendix A Figure 52: Group mean amygdala—mPFC BSC across contrasts, mPFC ROIs, and pipelines

Points represent mean estimates and error bars represent 95% confidence intervals.

Multiverse analyses of age-related change in amygdala—mPFC BSC

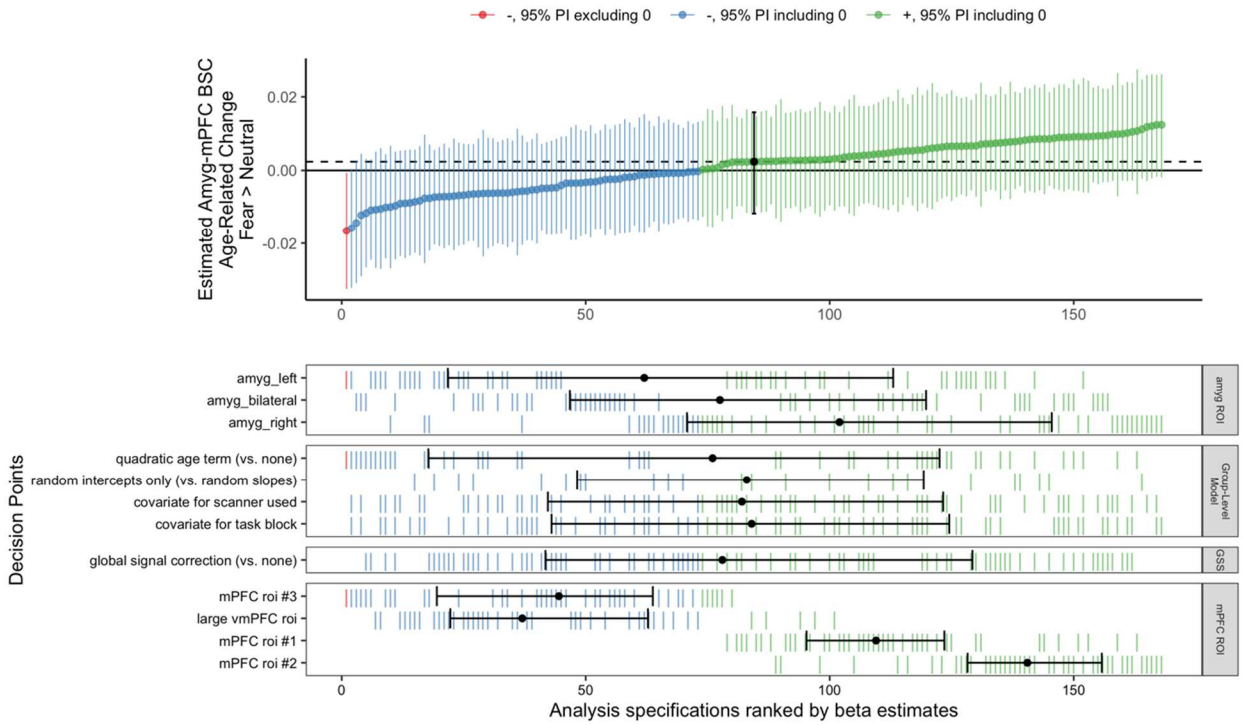
In addition to the constructing specification curves for age-related change in amygdala—mPFC BSC for the fear > baseline contrast as reported in the main manuscript (see Figure 5), we constructed parallel specification curves for the neutral > baseline and fear > neutral contrasts. For the neutral > baseline contrast, 77.9% of specifications found positive age-related change (most of them for mPFC ROI #1 or the large vmPFC ROI), though only 12.5% of posterior estimates excluded 0 (Appendix A Figure 53). For the fear > neutral contrast, only 1 pipeline out

of 300 resulted in a posterior estimate excluding 0 for age-related change (Appendix A Figure 54).



Appendix A Figure 53: Spec. curve for age-related change in neutral > baseline BSC

A: Points represent estimated linear age-related change in amygdala—mPFC BSC for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.



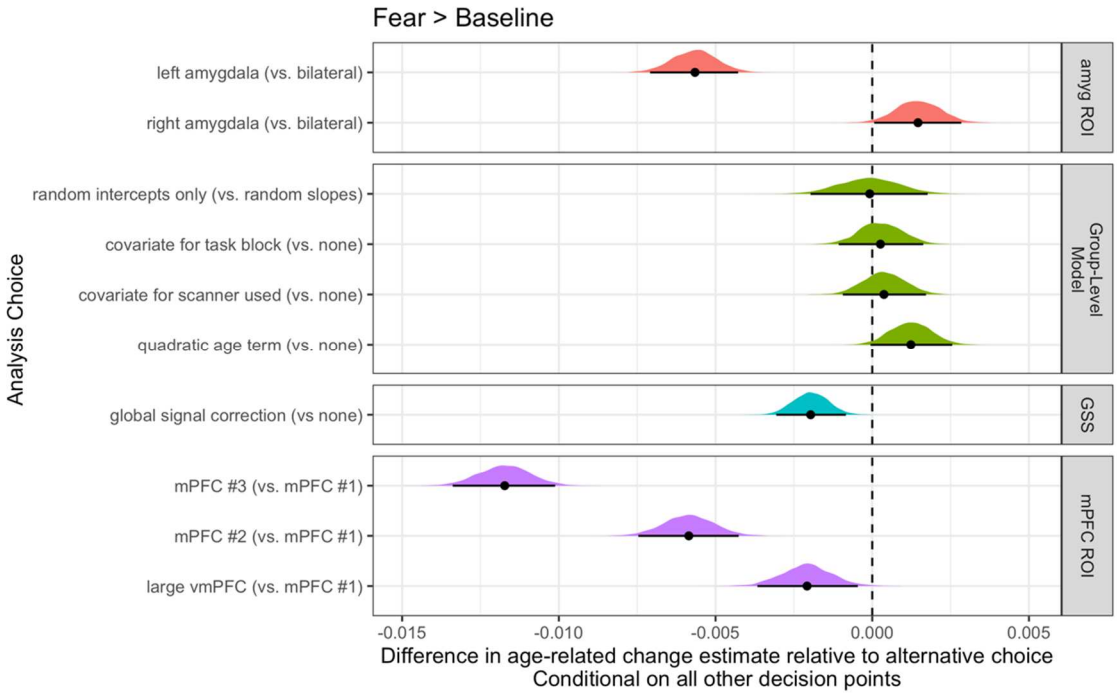
Appendix A Figure 54: Spec. curve for age-related change in fear > neutral BSC

A: Points represent estimated linear age-related change in amygdala—mPFC BSC for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.

Impacts of analysis choices on age-related change estimates for amygdala—mPFC BSC

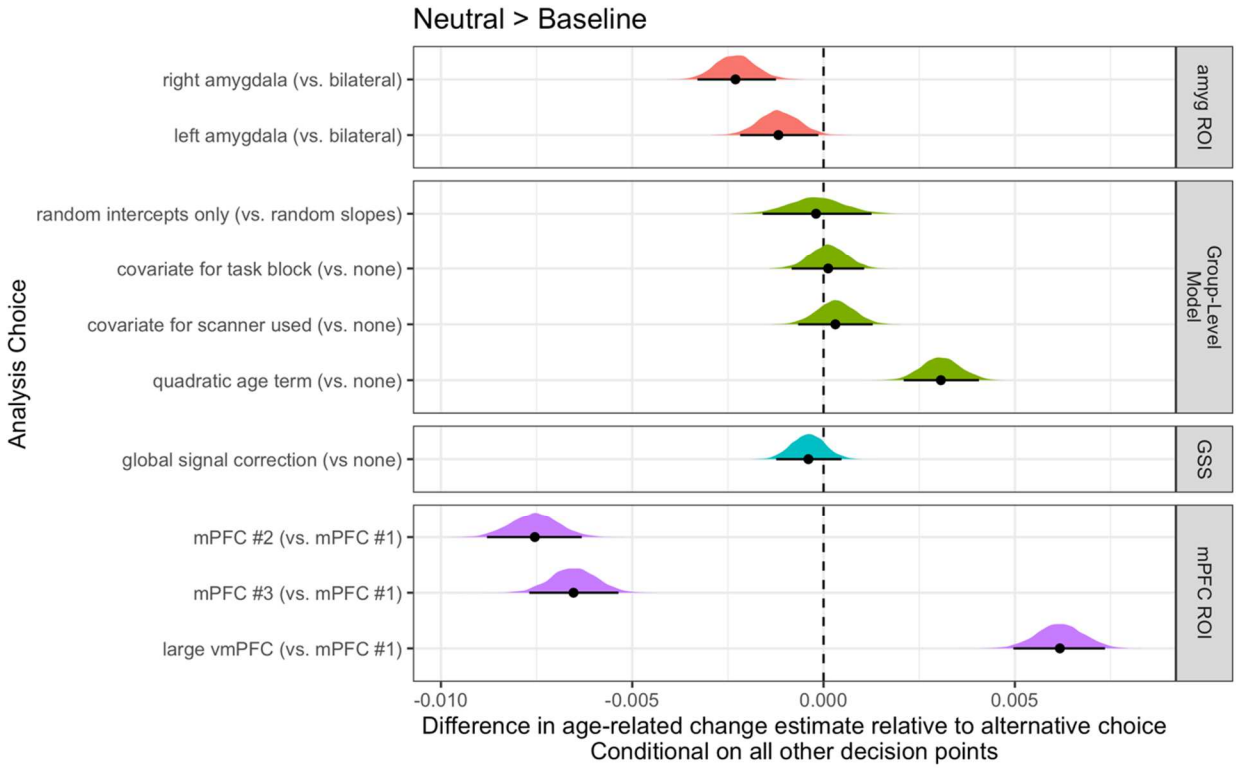
Unlike gPPI, for all BSC contrasts the choice of mPFC ROI, rather than preprocessing or modeling decisions, made biggest relative impact on estimates of age-related change (Appendix A Figures 55-57). For both fear > baseline and neutral > baseline contrasts, pipelines with mPFC ROIs #2-3 showed more negative age-related change compared to mPFC ROI #1 or the large vmPFC ROI. For the fear > neutral contrast, pipelines with mPFC ROI #3 and the large vmPFC

ROI showed more negative age-related change compared to mPFC ROIs #1-2 (Appendix A Figure 57).



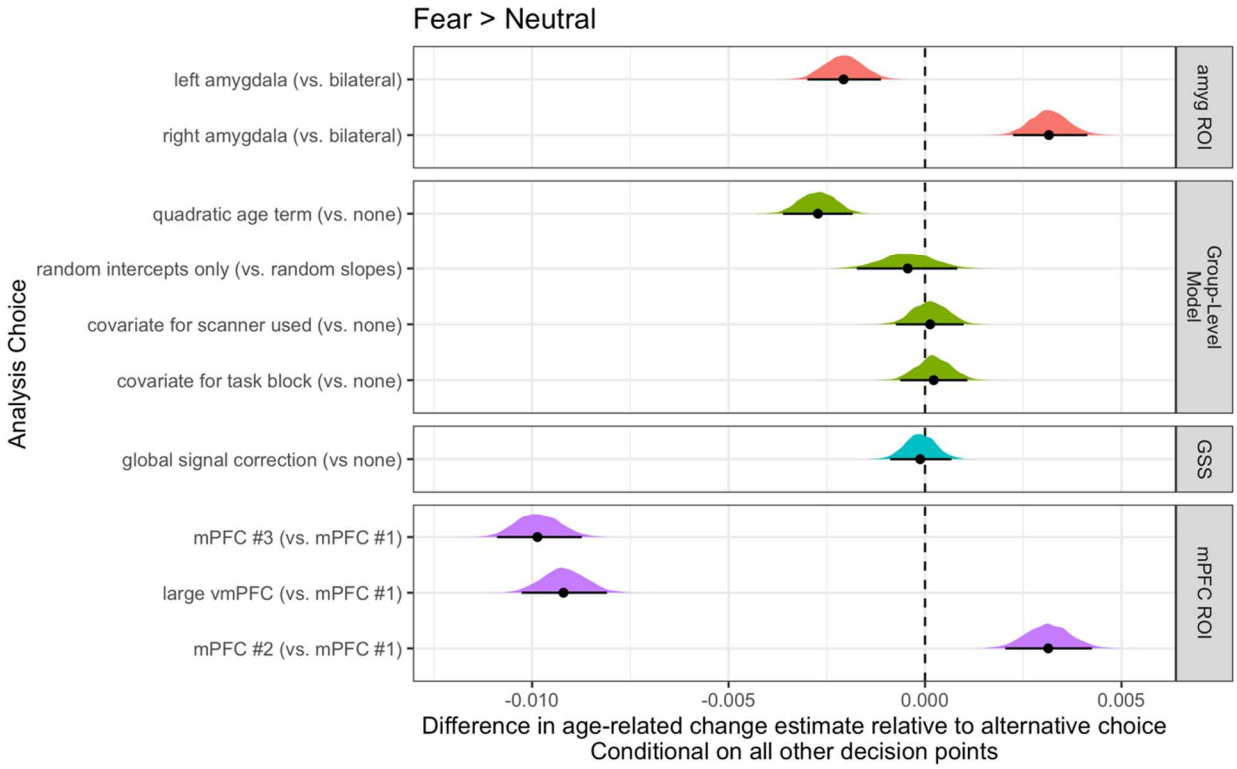
Appendix A Figure 55: Fork impacts on age-related change for fear > baseline amygdala—mPFC BSC

Posterior distributions and 95% posterior intervals are shown, representing the average difference in linear age-related change estimates relative to the alternative choice.



Appendix A Figure 56: Fork impacts on age-related change for neutral > baseline amygdala—mPFC BSC

Posterior distributions and 95% posterior intervals are shown, representing the average difference in linear age-related change estimates relative to the alternative choice.



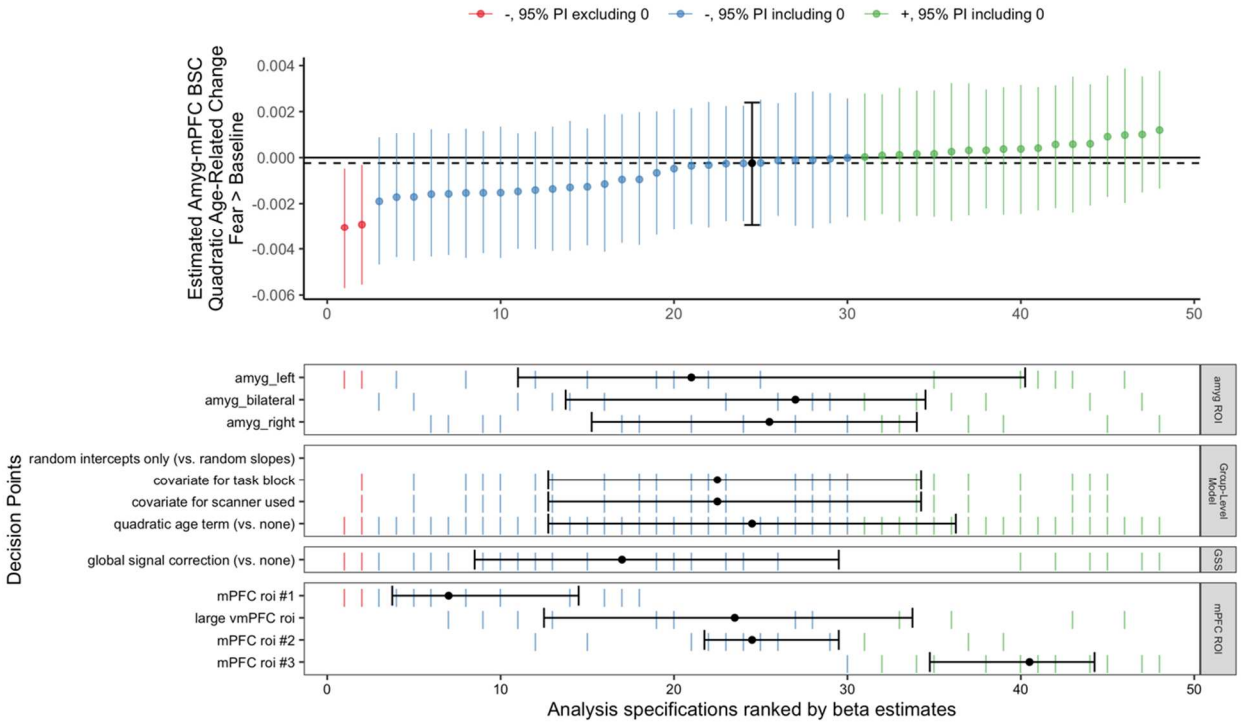
Appendix A Figure 57. Fork impacts on age-related change for fear > neutral amygdala—mPFC BSC

Posterior distributions and 95% posterior intervals are shown, representing the average difference in linear age-related change estimates relative to the alternative choice.

Nonlinear age-related changes in amygdala—mPFC BSC

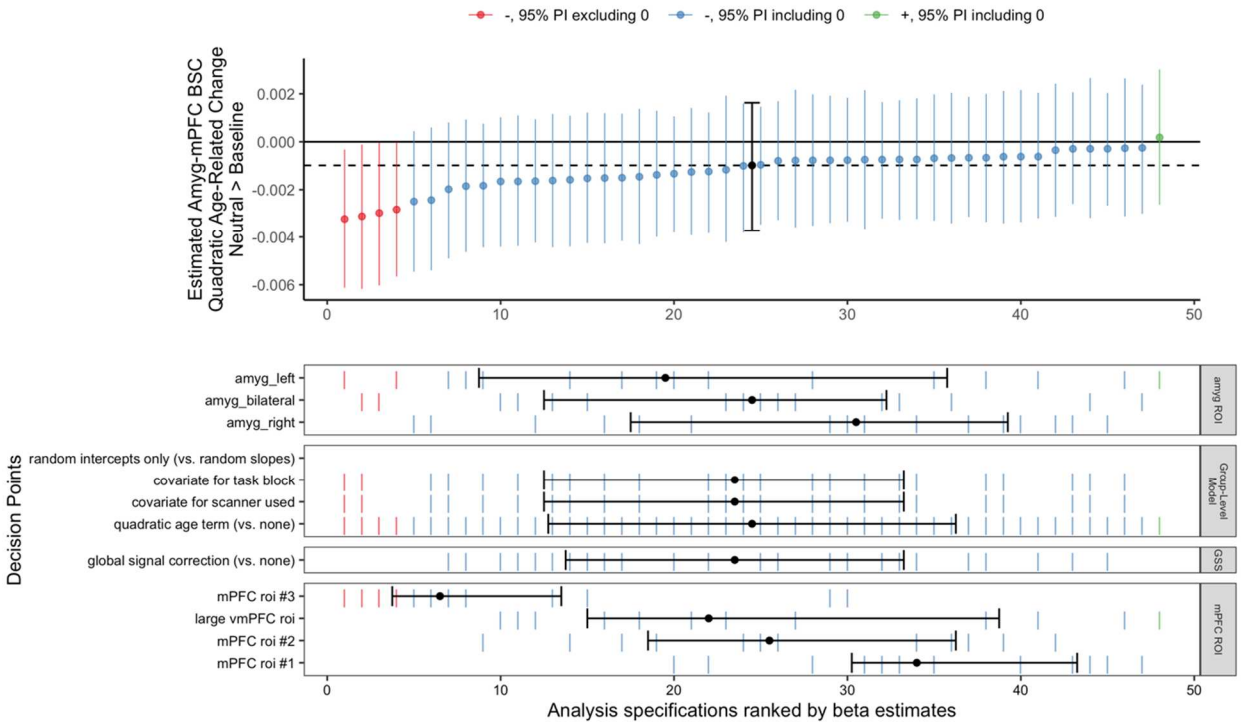
Some model specifications for amygdala—mPFC BSC included quadratic age-related change terms. Overall, however, we found little consistent evidence for quadratic age-related change for any contrast (see Appendix A Figures 58-60), as the sign of quadratic terms ('peaks' vs 'troughs') varied across specifications, and very few terms for each contrast were estimated with the 95% posterior interval excluding 0 (4% for fear > baseline, 8% for neutral > baseline, 18%

for fear > neutral). Inverse age models for age-related change in amygdala—mPFC BSC also indicated little evidence for consistent age-related change (Appendix A Figure 61).



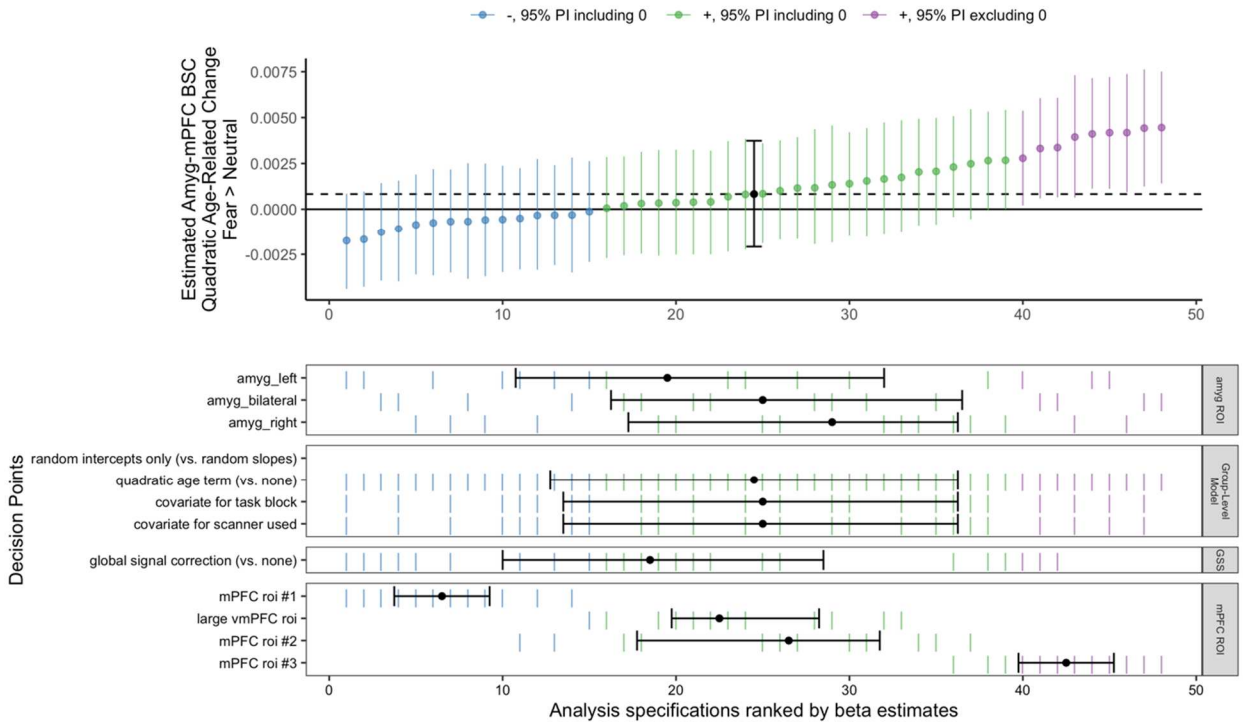
Appendix A Figure 58: Spec. curve for quadratic age-related changes in fear > baseline amygdala—mPFC BSC

A: Points represent estimated quadratic age-related change in amygdala—mPFC BSC for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.



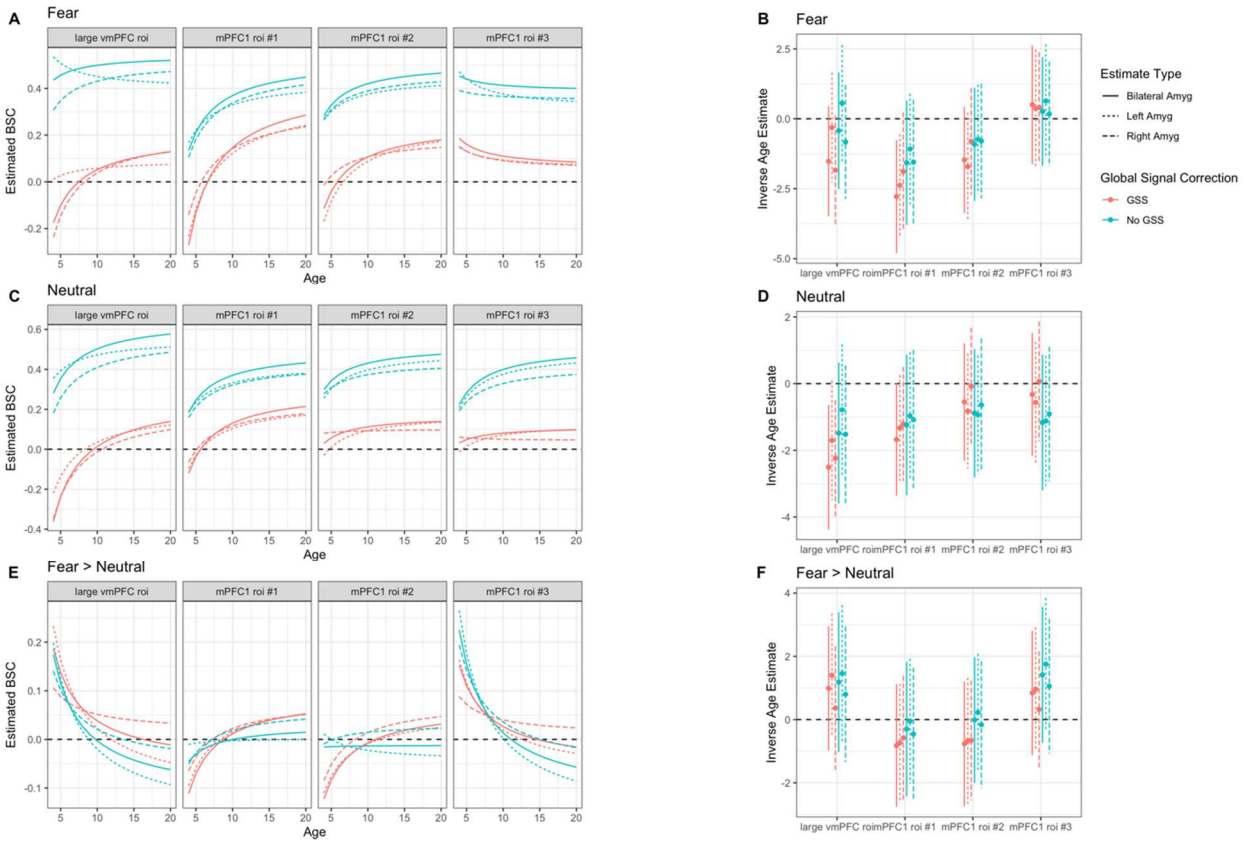
Appendix A Figure 59: Spec. curve for quadratic age-related changes in neutral>baseline amygdala—mPFC BSC

A: Points represent estimated quadratic age-related change in amygdala—mPFC BSC for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.



Appendix A Figure 60: Spec. curve for quadratic age-related changes in fear > neutral amygdala—mPFC BSC

A: Points represent estimated quadratic age-related change in amygdala—mPFC BSC for each specification, and lines represent corresponding 95% posterior intervals. B: Variables on the y-axis represent analysis choices, corresponding lines indicate that a choice was made, and blank space indicates that the choice was not made in a given analysis.



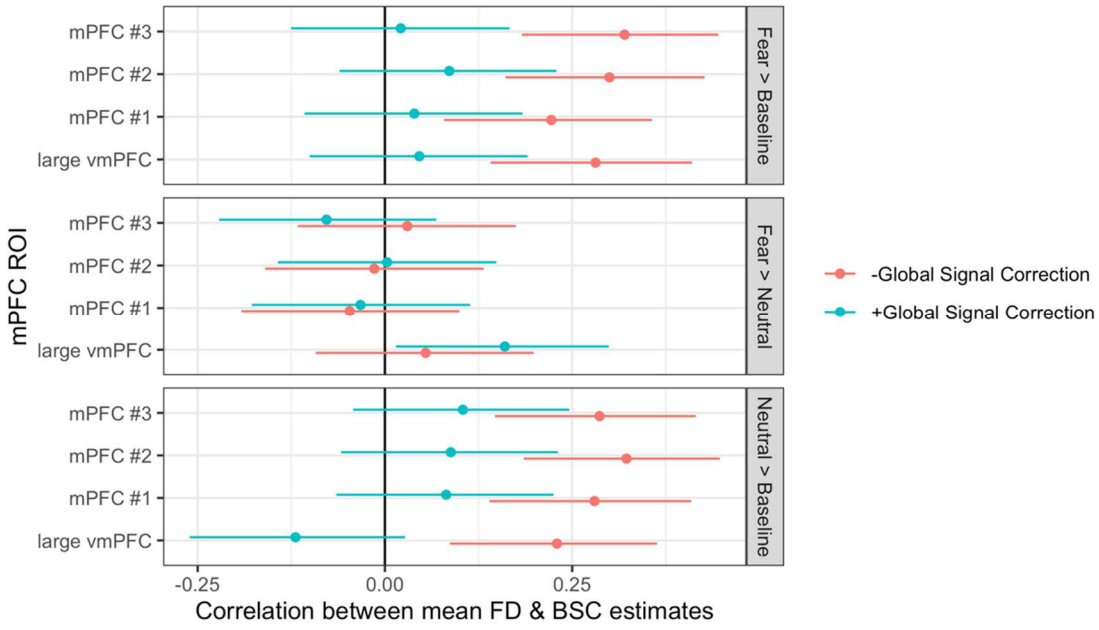
Appendix A Figure 61: Inverse age models for amygdala—mPFC BSC

Left panels show fitted model predictions for inverse age models for the fear (A, top), neutral (C, middle), and fear > neutral (E, bottom) contrasts. Specifications including a global signal correction are plotted in red, and without it in blue. Line type indicates amygdala ROI (solid = bilateral, dotted = left, dashed = right). Right panels show beta estimates for corresponding models for each contrast. Positive estimates for inverse age indicate decreases in amygdala—mPFC BSC as a function of age, and vice-versa.

Correlations between head motion & amygdala—mPFC BSC estimates

We calculated correlations between mean framewise displacement and BSC estimates for each ROI and contrast for pipelines with and without global signal correction. For pipelines

without global signal correction fear > baseline and neutral > baseline estimates were positively associated with head motion (see Appendix A Figure 62). Such correlations were reduced in pipelines with a global signal correction, consistent with indications that such estimates of BSC for only one condition may also represent some 'task-independent' signal that may contain motion and respiratory artifacts. BSC estimates for the fear > neutral contrast were not overall strongly associated with head motion.

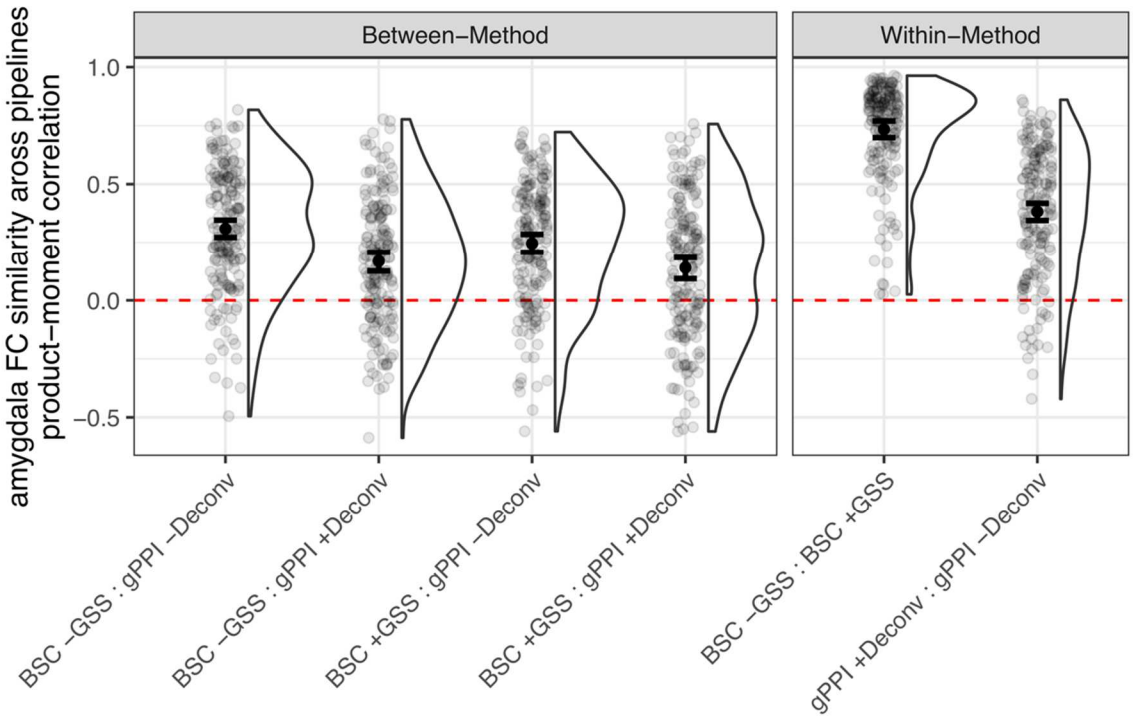


Appendix A Figure 62: Correlations between mean FD & BSC estimates across scans
For each contrast and ROI, points show estimated product-moment correlations and error bars represent 95% confidence intervals.

Within-scan similarity between gPPI & BSC amygdala FC

Previous work from Di et al. (2020) indicated a convergence in BSC and gPPI for contrasts between two task conditions. We addressed this in the present data with a within-scan analysis.To examine whether certain pipelines across BSC and gPPI were representing more

similar signals, we computed correlations between vectors of fear > neutral amygdala FC with the rest of the brain between each pair of BSC and gPPI pipelines for the same scan. Overall, patterns of amygdala FC with the rest of the brain were positively associated between all pipelines on average, although not strongly, and associations varied widely across participants (Appendix A Figure 63, left panel). gPPI pipelines without deconvolution resulted in somewhat more similar amygdala connectivity patterns with BSC relative to gPPI pipelines with deconvolution. Amygdala connectivity patterns were the most strongly associated between the two BSC pipelines with versus without global signal correction (Appendix A Figure 63, right panel).

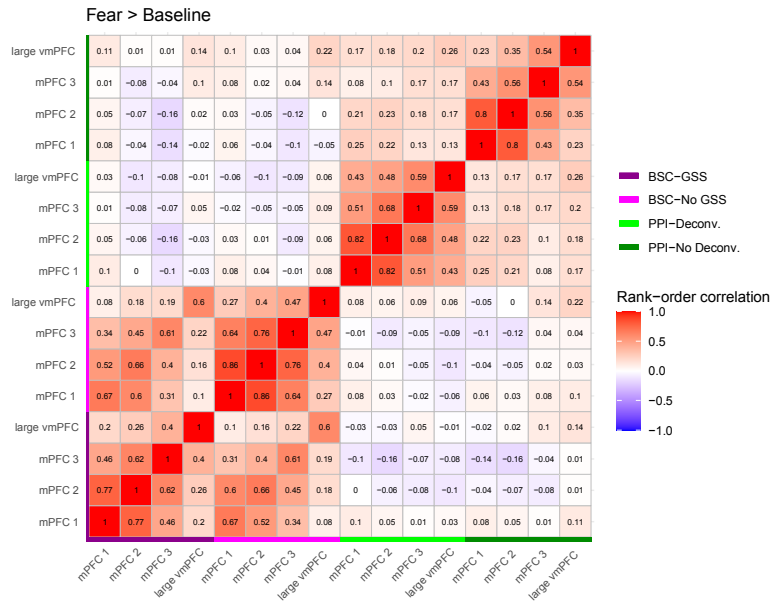


Appendix A Figure 63. Similarity of amygdala FC with the rest of the brain between gPPI & BSC

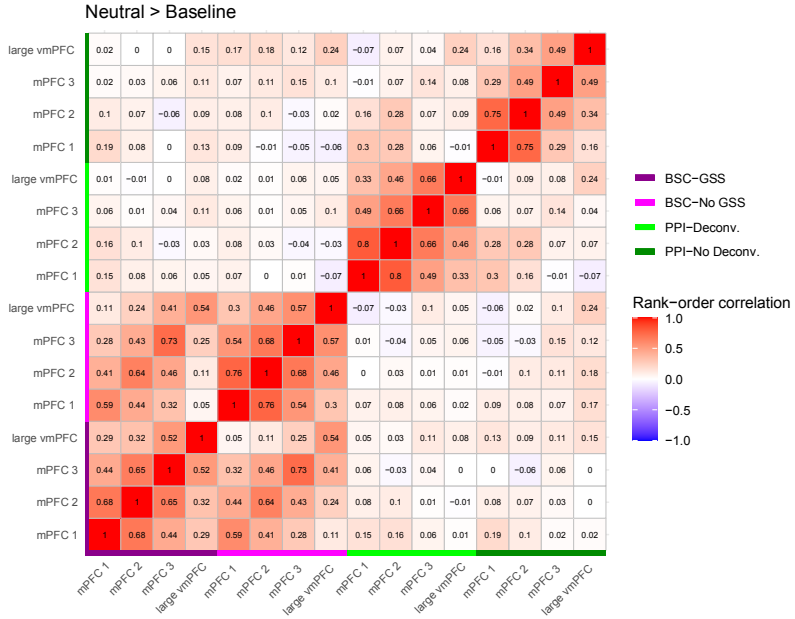
Points represent similarity across pipelines for each individual scan for each comparison (x axis marks each comparison of pipelines), with error bars summarizing 95% posterior intervals for estimated mean similarity across scans. Left: between-method comparisons of similarity of amygdala FC between BSC and gPPI methods. Right: within-method comparisons of similarity of amygdala FC within BSC and gPPI methods, respectively, while altering decisions for GSS (for BSC) and deconvolution (gPPI). GSS = global signal subtraction using post-hoc mean centering.

Between-scan correlations between gPPI & BSC estimates

We also sought to examine whether different methods for functional connectivity yielded similar relationships between scans in mean amygdala—mPFC functional connectivity estimates. We computed between-scan rank-order correlations between gPPI and BSC estimates for each contrast to examine whether relative ordering was preserved across different estimates of amygdala—mPFC functional connectivity. For each pair of preprocessing pipelines, we computed the rank-order correlation between vectors of functional connectivity estimates (1 datapoint per scan per pipeline). Most generally, BSC estimates were not strongly associated with gPPI estimates, even for the same contrast and mPFC ROI (Appendix A Figures 64-66). However, for the fear > neutral contrast specifically, gPPI estimates without deconvolution were generally positively correlated with BSC estimates for corresponding ROIs, while gPPI estimates with deconvolution were not (Appendix A Figure 66). In addition, while BSC estimates for the same ROI with versus without global signal correction were generally highly associated, gPPI estimates for the same ROI with versus without deconvolution were often only weakly associated.

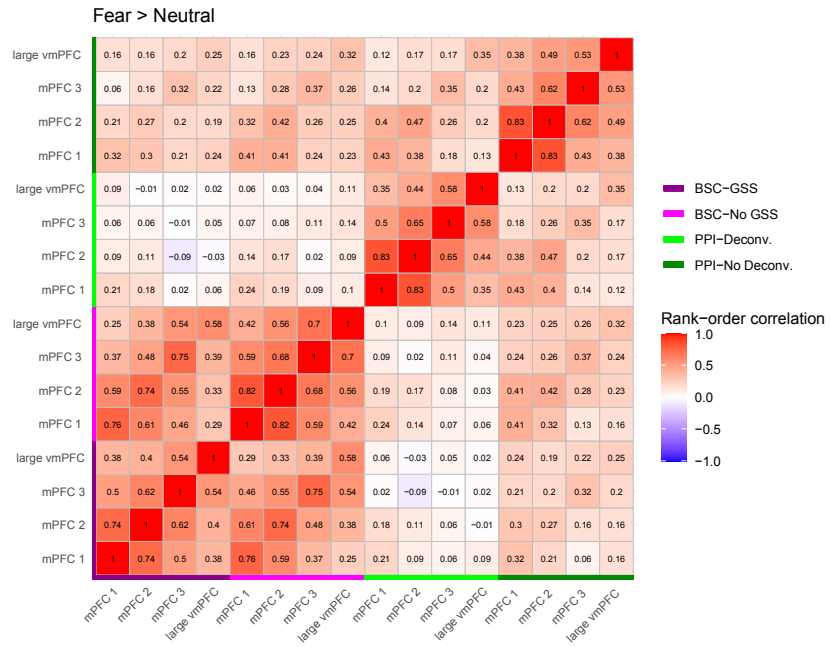


Appendix A Figure 64: Between-scan correlations between fear > baseline gPPI & BSC



estimates

Appendix A Figure 65: Between-scan correlations between neutral > baseline gPPI & BSC estimates

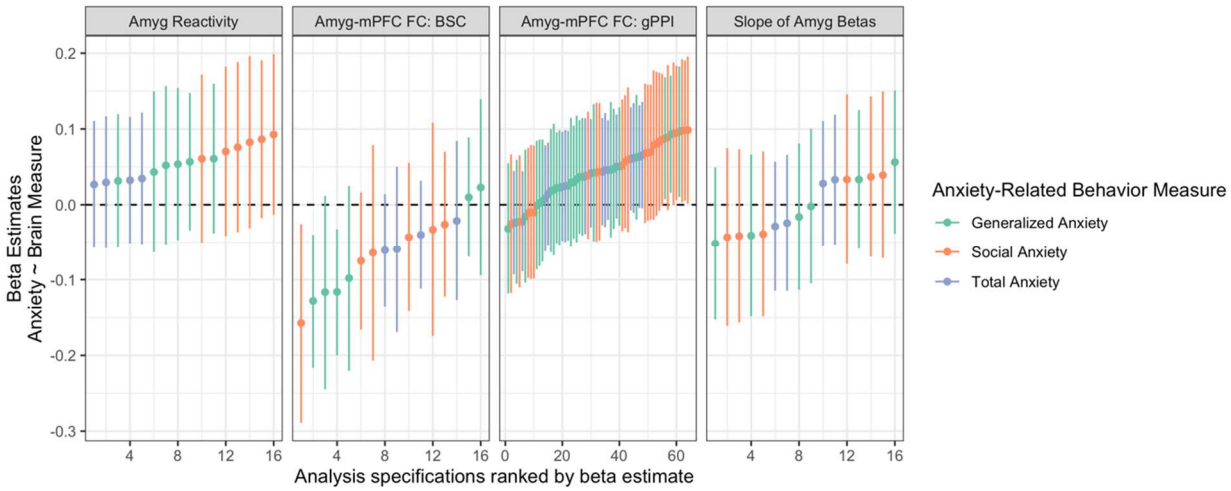


Appendix A Figure 66: Between-scan correlations between fear > neutral gPPI & BSC estimates

Associations with generalized anxiety and social anxiety behaviors

Here, we primarily focused on associations between amygdala—mPFC responses and separation anxiety in efforts to follow up on previous analyses of the same cohort (Gee et al., 2013), and because separation anxiety behaviors often show pronounced decline among typically developing cohorts of age range (Allen et al., 2010; Francis et al., 1987). However, we also examined associations with parent-reported generalized anxiety and social anxiety behaviors as measured by the SCARED-P and RCADS-P, as well as the Total Anxiety Score from the RCADS-P. Using multilevel linear regression models (covarying for age) with maximum

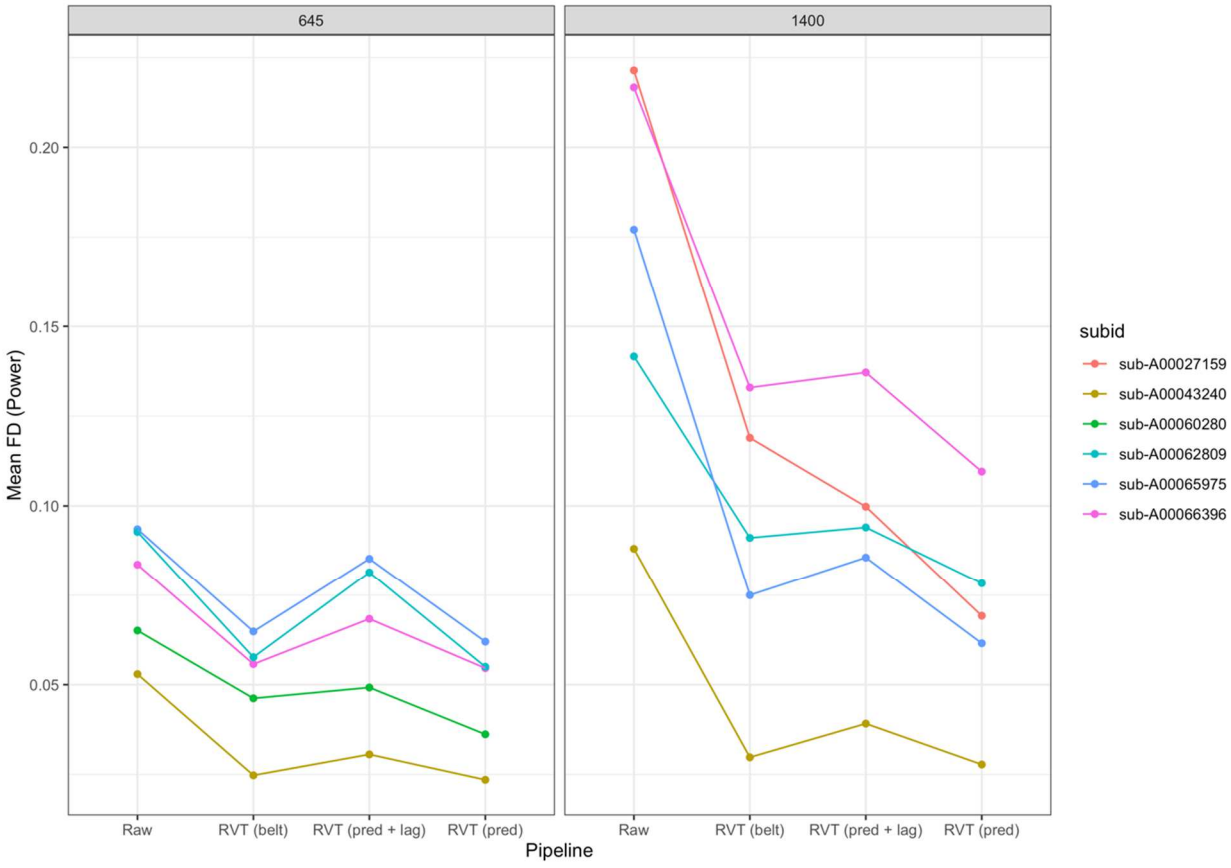
likelihood estimation as described above, we did not find robust evidence for longitudinal associations between amygdala—mPFC measures and any of the anxiety scales examined (Appendix A Figure 67).



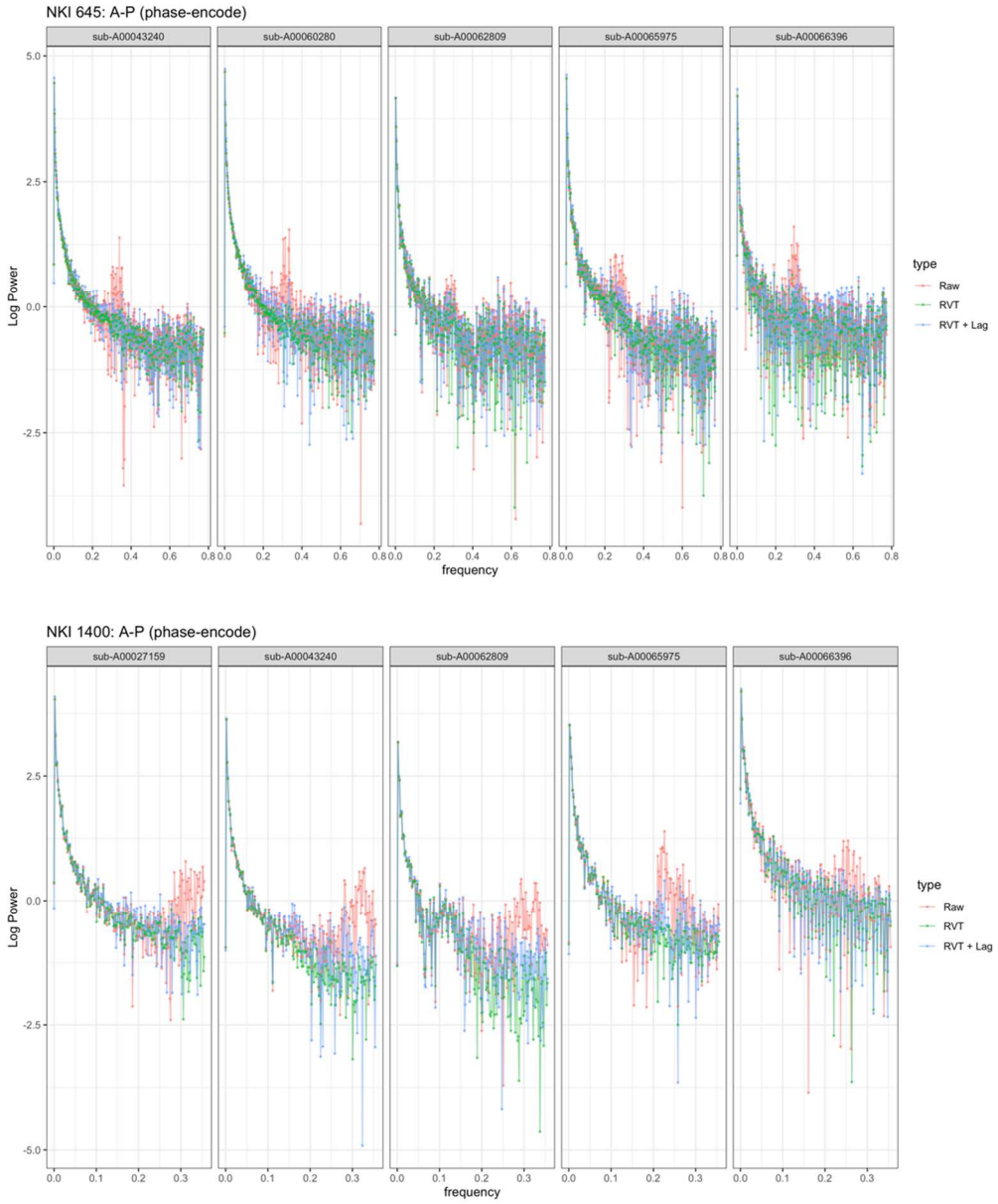
Appendix A Figure 67: Associations between amygdala—mPFC measures and anxiety-related behaviors

Specification curves are shown for longitudinal associations between fear > baseline amygdala reactivity (A), fear > baseline amygdala—mPFC BSC (B), fear > baseline amygdala—mPFC gPPI (C), and slopes for amygdala fear betas (D) and anxiety scales (green = generalized anxiety behaviors, orange = social anxiety behaviors, purple = total anxiety behaviors).

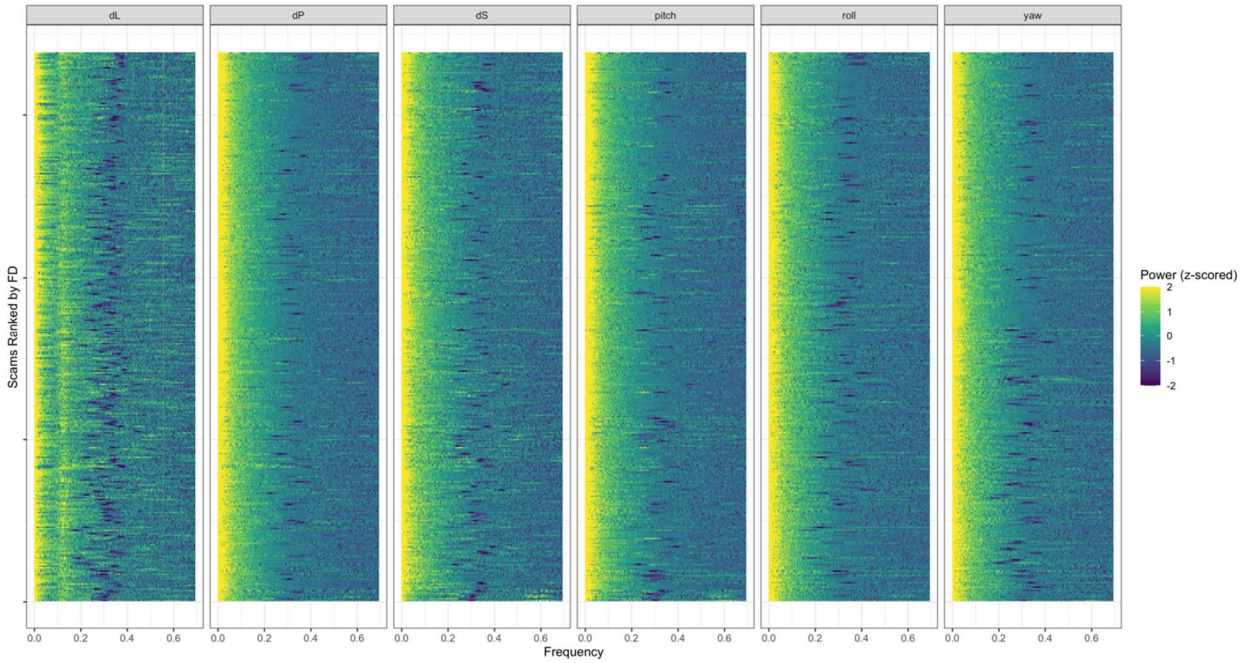
Appendix B: Chapter 2 Supplement



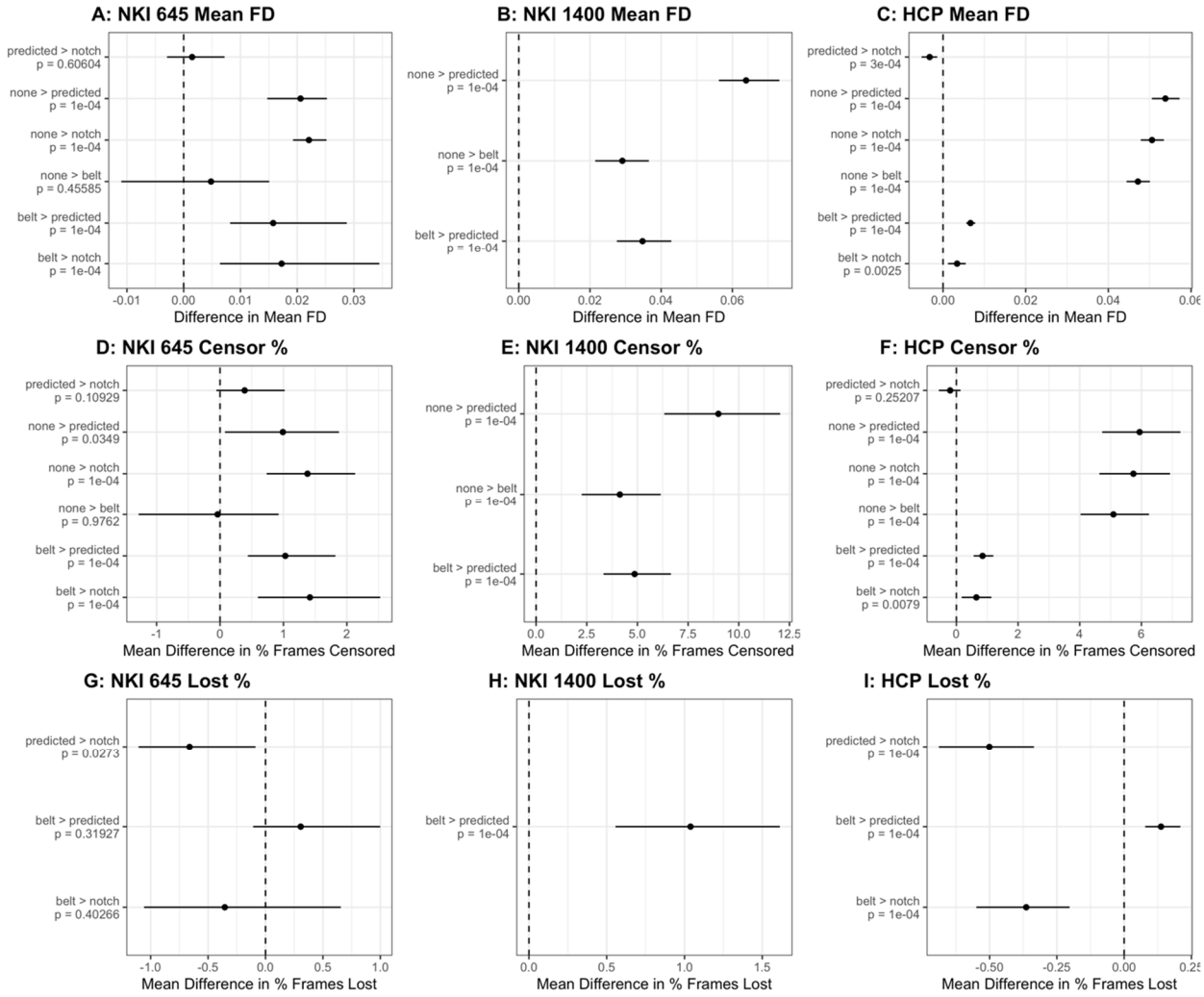
Appendix B Figure 1: Estimated head motion for a subset of NKI participants with lagged predicted RVT + RETROICOR compared to without lag, raw BOLD, and belt RVT + RETROICOR. Panels show TR=645 on the left and TR=1400 on the right.



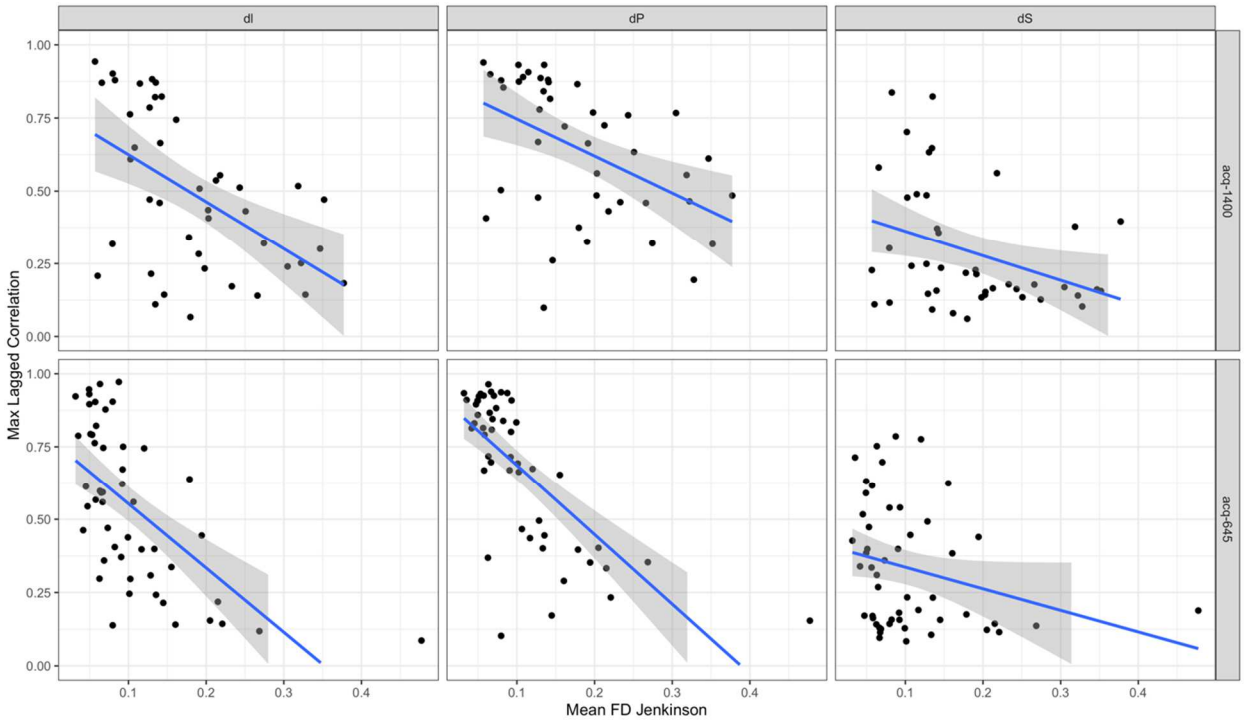
Appendix B Figure 2: Estimated power spectra of head motion in the anterior-posterior (phase-encoding) direction for a subset of NKI participants with lagged predicted RVT + RETROICOR compared to without lag, raw BOLD, and belt RVT + RETROICOR. Top panel shows TR=645 and bottom panel shows TR=1400.



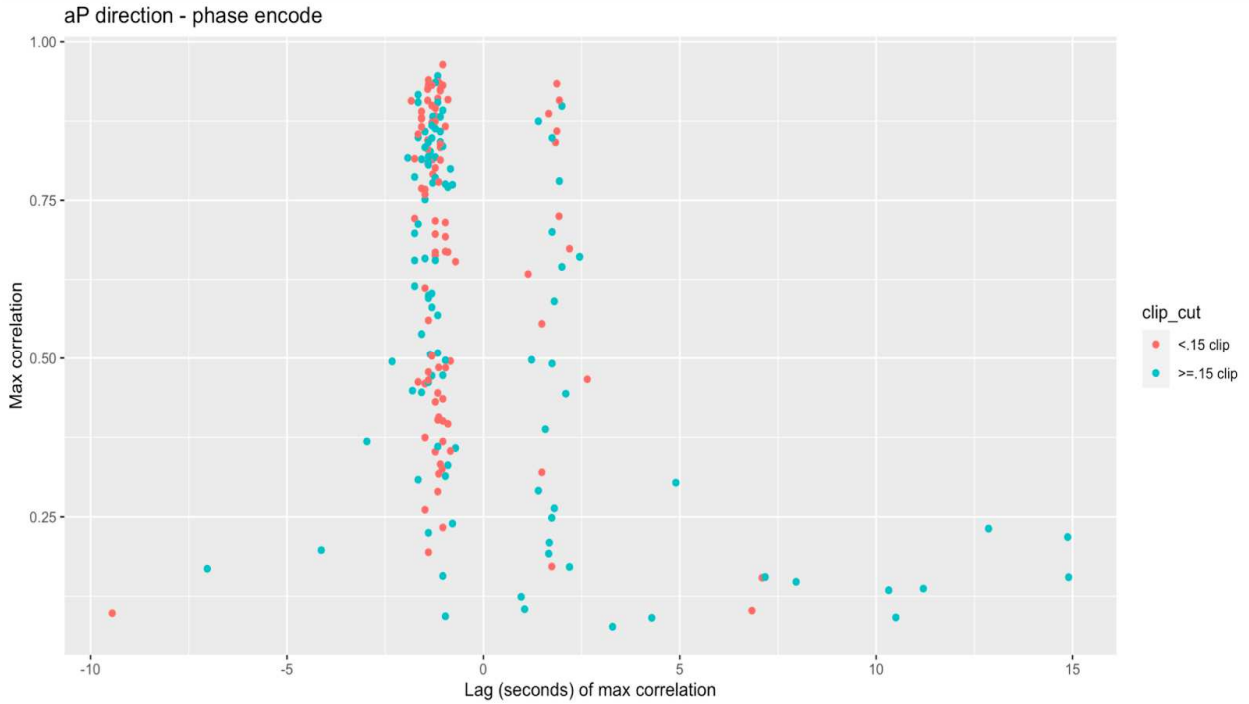
Appendix B Figure 3: Power spectra in the HCP data when an "adaptive notch" filter was applied. The notch filter center was set to the scan-specific predicted peak respiratory frequency with a width of 0.12Hz.



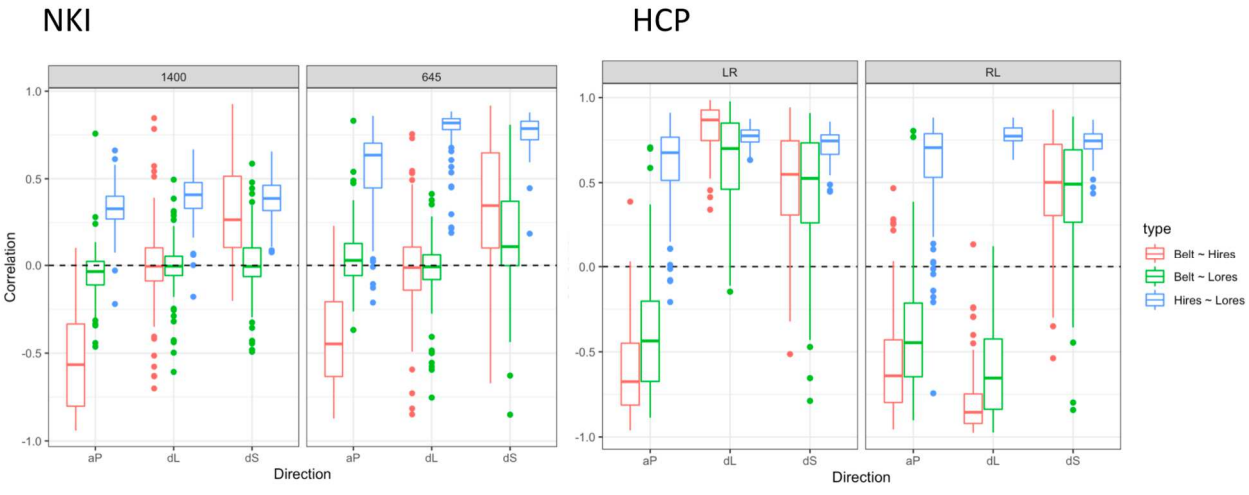
Appendix B Figure 4: Bootstrap comparisons for head motion metrics between raw BOLD, notch filtering, belt RVT + RETROICOR, and predicted RVT + RETROICOR. Each point range shows the bootstrapped mean and 95% confidence interval for the mean difference in each metric (top row = mean Jenkinson FD, middle row = % TRs censored, bottom = % TRs lost). Y-axis labels indicate the contrast.



Appendix B Figure 5: Relationships between head motion (x-axis) and maximum lagged correlations between belt and high-resolution motion timeseries within the NKI data. Columns show direction (left-right, anterior-posterior, inferior-superior), and rows indicate sequence. All combinations of sequence and direction indicate that maximum lagged correlations are weaker in scans with higher mean Jenkinson FD.

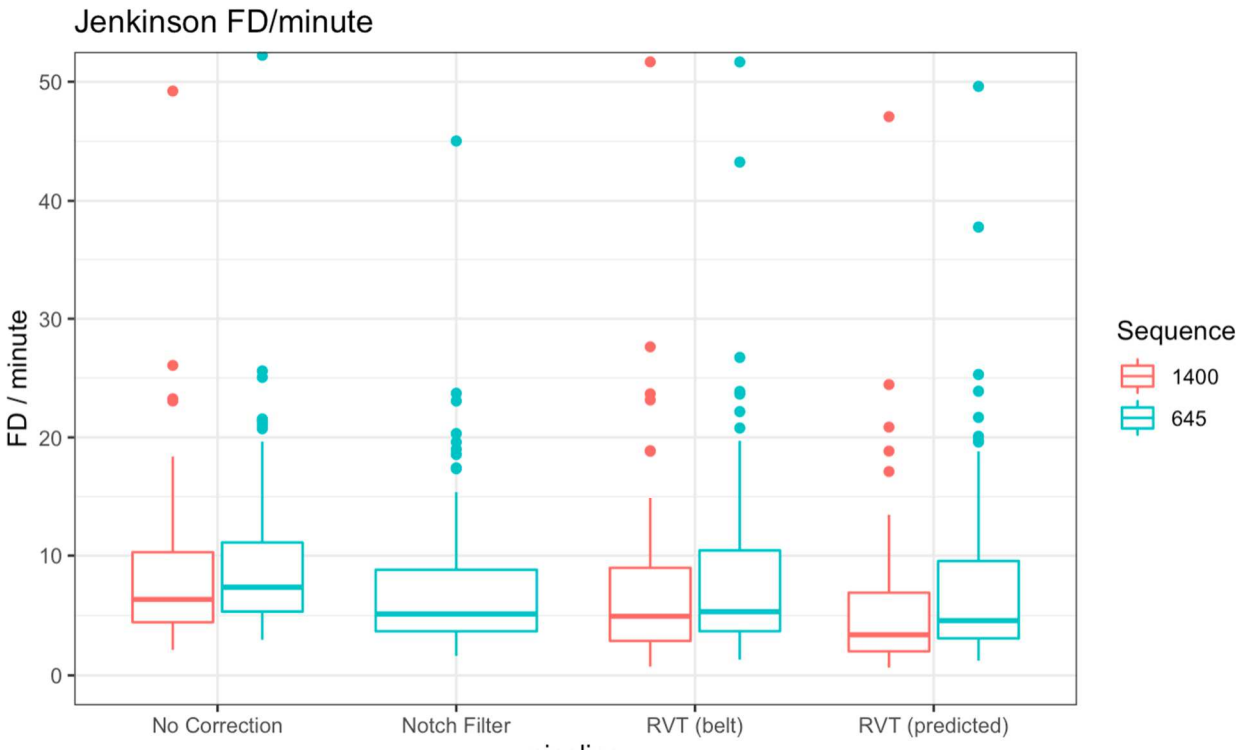


Appendix B Figure 6: Distributions of lag times for the maximum correlations between belt and predicted respiratory traces within the NKI data (A-P direction). X-axis indicates the temporal lag (positive indicates the predicted traces was "ahead") and y-axis indicates the max correlation found at the respective lag time. Blue dots indicate scans where >= 15% of belt trace time points were clipping, red dots indicate < 15% clipping.

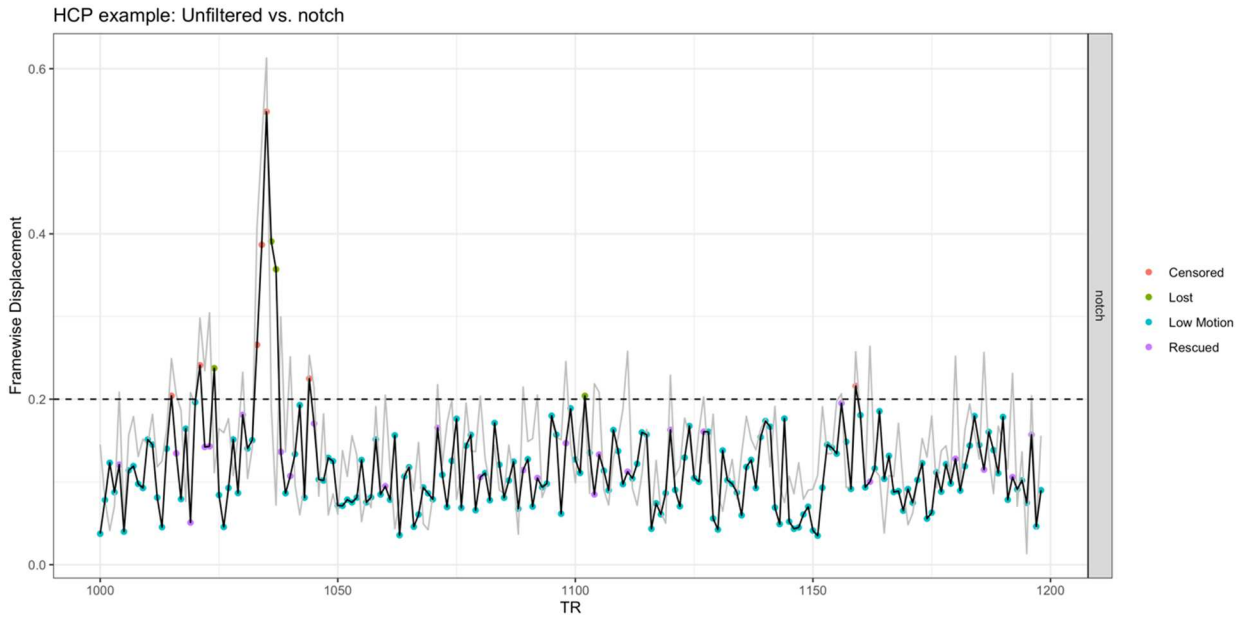


Appendix B Figure 7: Non-lagged correlations between the respiratory belt and high-resolution (red) and original resolution (green) motion parameters in the 0.2-0.6Hz range. Correlations are

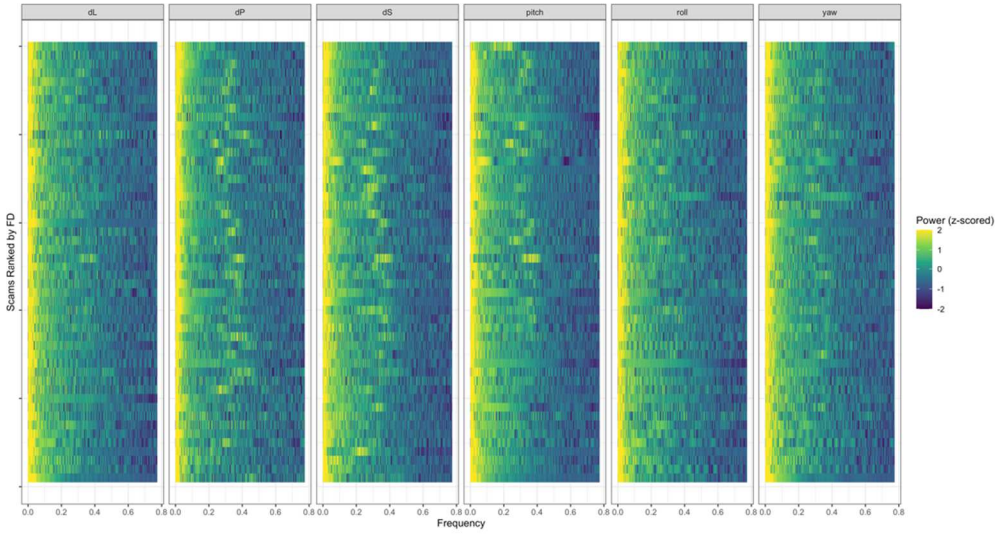
shown in each direction of translation, and also for relationships between high resolution and original resolution motion parameters (blue). Correlations are shown for the NKI (left) and HCP (right) data



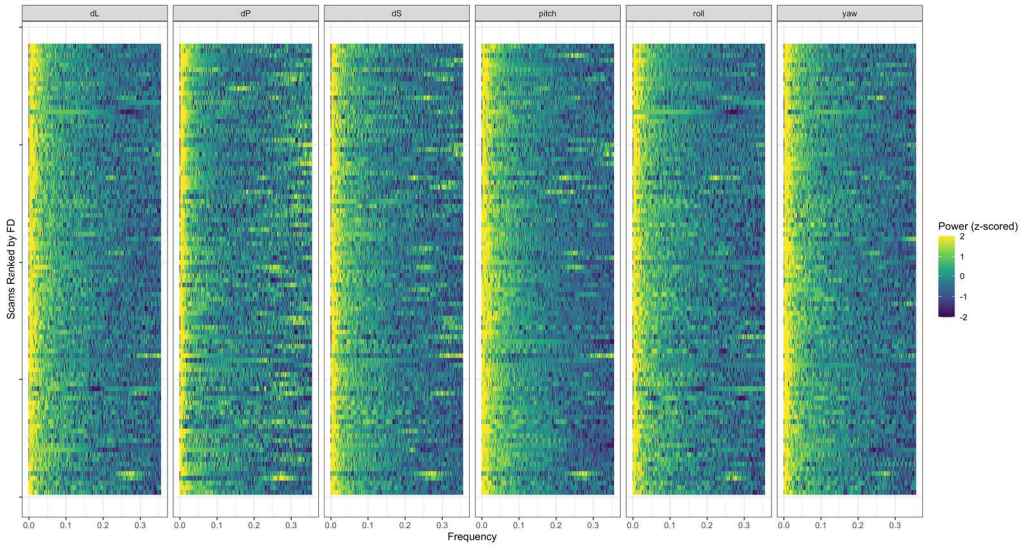
Appendix B Figure 8: Normalizing framewise displacement estimates to FD-per-minutes reduces discrepancies in head motion estimates across sequences.



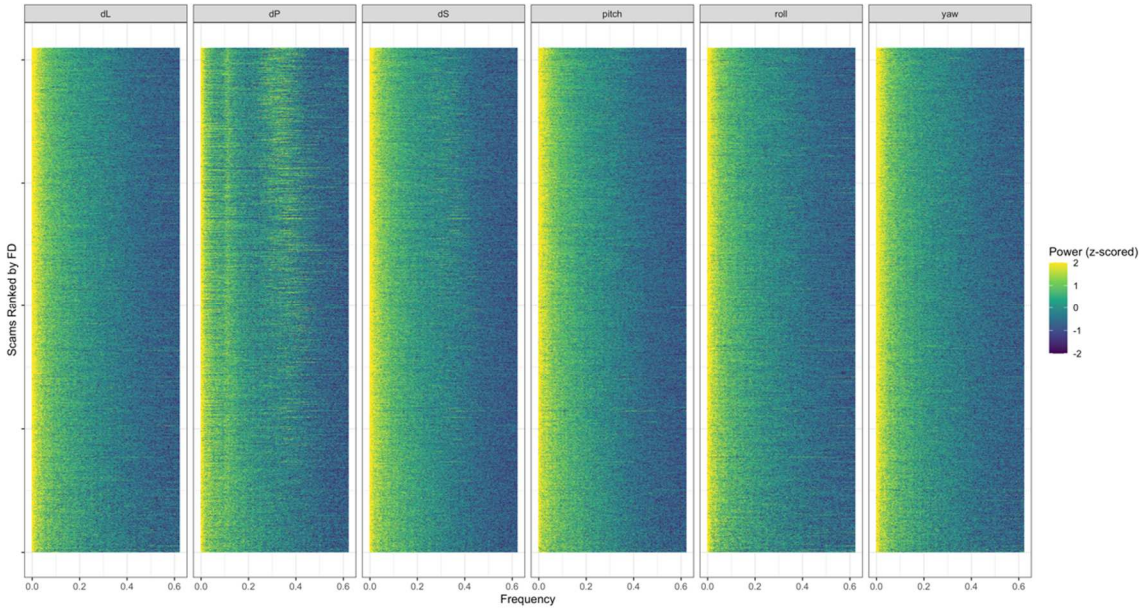
Appendix B Figure 9: Comparison of an example HCP participant's estimated FD from raw BOLD (light grey) and after notch filtering (black, with points). Lost TRs where only the notch filtered frame is above FD = 0.2mm occur after the large motion spike. Motion may thus be "spread" from large spikes to neighboring TRs.



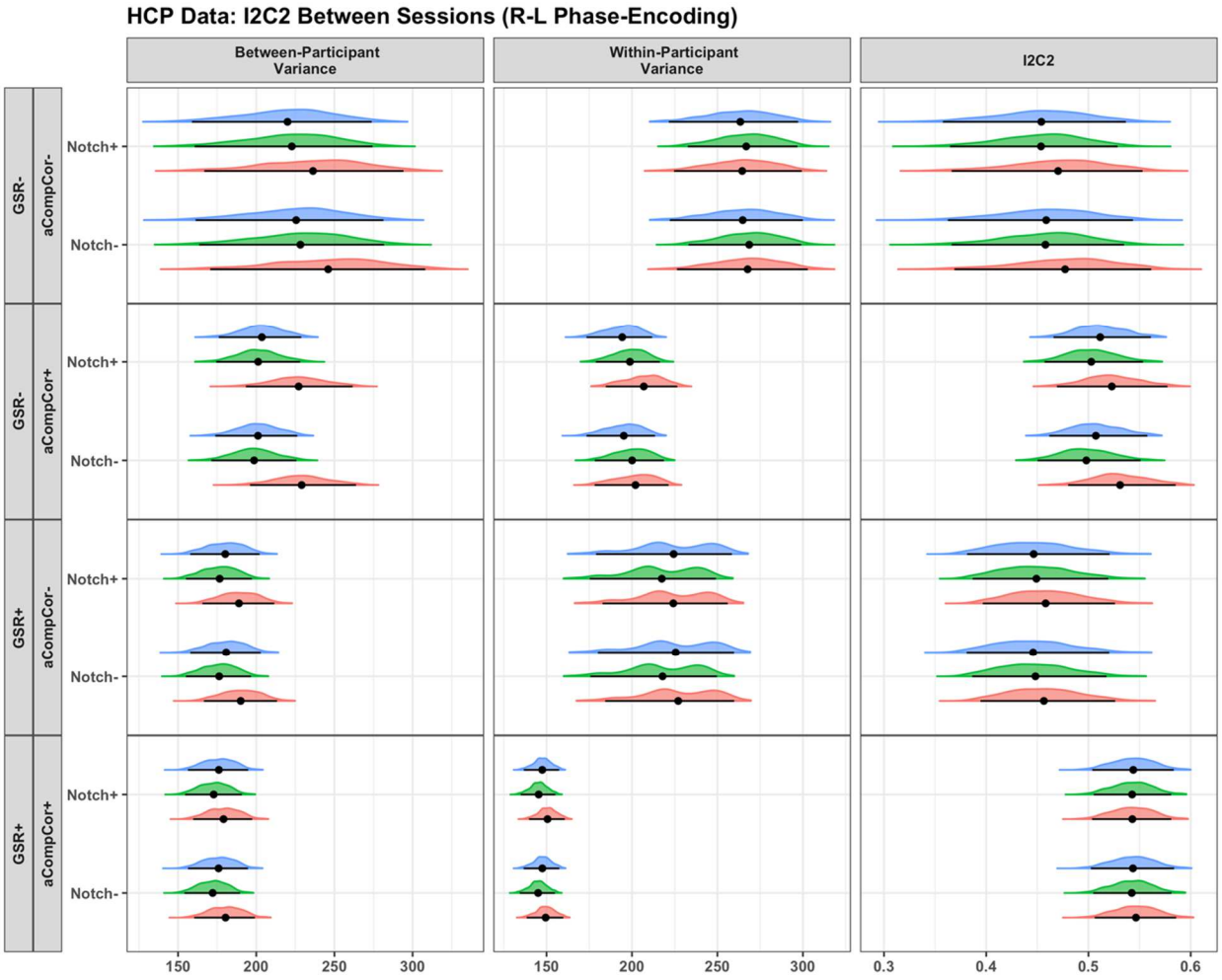
Appendix B Figure 10: Power spectra of head realignment parameters in the NKI 645 raw BOLD data



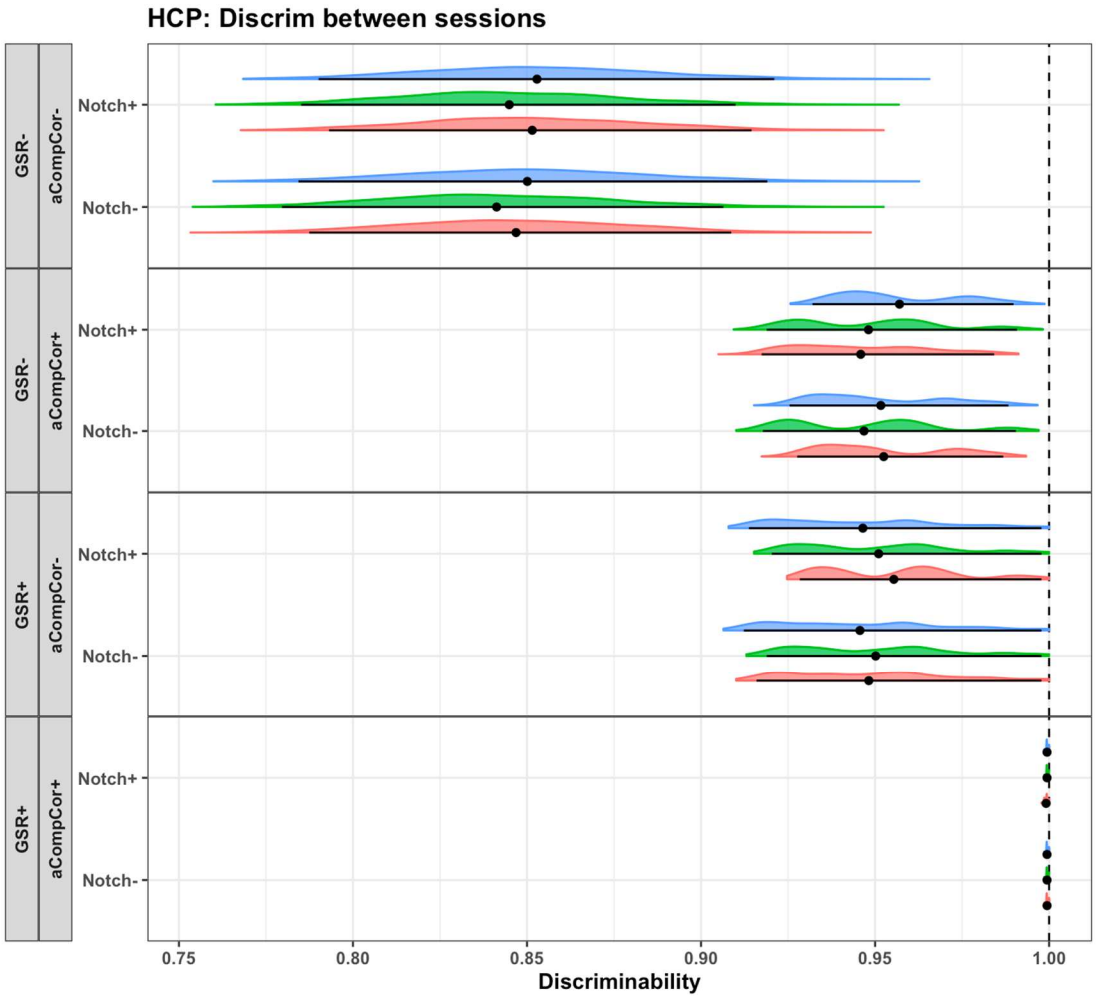
Appendix B Figure 11: Power spectra of head realignment parameters in the NKI 645 raw BOLD data



Appendix B Figure 12: Power spectra of head realignment parameters among resting-state scans collected as part of the PACCT study (Tottenham). Data were collected in the A-P phase encoding direction on a Siemens Prisma with multiband factor=6.



Appendix B Figure 13: Bootstrapped I2C2 within HCP data collected in the R-L phase encoding direction. In the R-L, compared to L-R data, aCompCor made somewhat of a clearer improvement to I2C2. Red = no model-based physio correction, green = belt RVT + RETROICOR, blue = predicted RVT + RETROICOR.



Appendix B Figure 14: Bootstrapped discriminability within HCP data collected in the R-L phase encoding direction. Red = no model-based physio correction, green = belt RVT + RETROICOR, blue = predicted RVT + RETROICOR.

Appendix C: Chapter 3 Supplement

Parent-selected child race	*N*	*Percentage*
European-American/Caucasian	23	37.1%
African American/Black	22	35.5%
Other	21	33.9%
Asian American	2	3.2%
American Indian/Alaska Native	1	1.6%
Native Hawaiian/Other Pacific Islander	0	0%

Appendix C Table 1: Parent-selected child race. Parent's selected from a list of options whether their child was of each race or not. Parents could select multiple options; therefore total percentages add up to over 100%. 20 parents (32.2%,

Parent-reported child race	*N*
Hispanic	6
Hispanic/Caucasian	2
Italian	2
White	2

American/Hispanic	1
Arabic African	1
Caribbean	1
Caucasian+Asian	1
East Indian	1
Eurasian	1
Hispanic-American	1
Latino	1
White & Asian	1

Appendix C Table 2: Parent-reported child race for parents who initially reported their child's race as "other". In this case, responses are shown as given by parents.