Unintended bias in the pursuit of collinearity solutions in fMRI analysis

Jeanette A. Mumford[1], Michael I. Demidenko[1], James M. Bjork[2], Hugh P. Garavan[3], Donald J. Hagler Jr.[4], Steven M. Nelson[5], Tor D. Wager[6], and Russell A. Poldrack[1]

[1]Department of Psychology, Stanford University

[2]Institute for Drug and Alcohol Studies, Virginia Commonwealth University

[3]Department of Psychiatry, University of Vermont

[4]Department of Radiology, Center for Multimodal Imaging and Genetics, University of California, San Diego

[5]Department of Pediatrics, Masonic Institute for the Developing Brain University of Minnesota

[6]Department of Psychology and Neuroscience, Dartmouth College

Abstract

In task functional magnetic resonance imaging (fMRI), collinearity between task regressors in a design matrix may impact power. Researchers often optimize task designs by assessing collinearity between task regressors to minimize downstream effects. However, some methods intended to reduce collinearity during optimization and data analysis may fail, in some cases introducing unintended bias into the parameter estimates. Although relevant to all task-based fMRI studies, we describe these issues and illustrate them using the Monetary Incentive Delay (MID) task fMRI data from the Adolescent Brain Cognitive Development (ABCD®) study. Specifically, we show that omitting regressors for certain task components, using impulse regressors for extended activations, and ignoring response time adjustments can bias common contrast estimates. We present a "Saturated" model that models all stimuli and response times, which minimizes bias in MID task simulations and validly estimates task-relevant whole brain activity, offering greater flexibility in studying contrasts that might otherwise be avoided due to potential biases.

Introduction

Functional magnetic resonance imaging (fMRI) tasks often produce signals that are challenging to detect, especially when studying individual differences [Poldrack et al., 2017, Elliott et al., 2020, Demidenko et al., 2024]. To address this, researchers prioritize models and study designs that maximize statistical power. Since collinearity between regressors can reduce power, efficiency optimization is often used prior to data collection to minimize collinearity [Liu et al., 2001]. We assume the models used in design optimization and data analysis are identical to ensure the optimization results align as closely as possible to real data outcomes, aside from differences introduced by behavior-based or motion regressors that are unknown during design optimization. Certain tasks, like the Monetary Incentive Delay (MID) task, are inherently at risk for high collinearity. This arises because stimuli that drive distinct neural processes are presented in rapid succession without intervening resting fixation. As a result, the BOLD signals associated with these stimuli overlap significantly, leaving minimal unique, non-overlapping signal to support the estimation of activation due to each stimulus separately. Here we will focus on the version of this task used in the Adolescent Brain Cognitive Development (ABCD®) study, in which each trial consists of a cue, fixation, probe, and feedback presented consecutively, with no intervals in between. This type of design is likely to have substantial overlap between the signals emerging from adjacent task components within a trial, and a common practice to resolve collinearity is to omit regressors for a subset of trial components (e.g., fixation and probe in the MID task). However, this common practice can bias the remaining estimates, making any analysis of "power benefit" due to collinearity reduction irrelevant. Here, we examine these biases using the MID task data from the ABCD Study. We illustrate the unintentional biases that occur as a result of the omissions and propose a modeling strategy to eliminate bias in the analysis of MID and other similarly designed fMRI tasks with closely spaced trial components.

This paper is divided into four parts. First, we review best practices for modeling fMRI data, focusing on how to assess collinearity, when it is acceptable and how to address it. Second, we introduce the MID task, its history, its design as implemented in the ABCD® study and a brief overview of historical modeling decisions. Third, we demonstrate how modeling choices affect key contrast estimates for the MID task, using simulations to highlight potential biases from unmodeled stimuli and then apply these analyses to real fMRI data from the ABCD® study's MID task acquisition. Finally, we discuss our findings and provide recommendations. Although we are focusing on the modeling strategy of the MID task, our results generalize to task fMRI modeling, in general.

fMRI time series modeling and collinearity

A common error when constructing linear regression models is to omit relevant variables, which can introduce bias [Greene, 2003, Chapter 8]. To avoid the omission error when modeling fMRI time series data, it is essential to model all presented stimuli. To ensure good fit to the data, this includes using the true duration of the neural responses, and to account for response time signals [Mumford et al., 2024]. If a stimulus is omitted from a model, the signal from that stimulus may be absorbed into activation estimates of other stimuli, resulting in bias and invalidating statistical inferences. Many fMRI tasks include baseline stimuli (e.g., resting fixation or intertrial intervals), which may be left unmodeled unless there is evidence that the neural response is altered in a relevant way during that period. As we will explain shortly, the fixation period during the MID task is not a true period of rest and, therefore, should be modeled. Collinearity between regressors is inevitable, due to the rapid timing of the trial components and the slower BOLD fMRI hemodynamic delay, and usually should be accepted without altering the model.

Collinearity can be minimized in advance by including an efficiency analysis and variance inflation factor (VIF) estimation during the study design phase. For a contrast, c, of a vector of parameter estimates, $\hat{\beta}$, the variance of the contrast estimate can be expressed as $Var(c\hat{\beta}) = \sigma^2/\text{eff}_{c\hat{\beta}}$, where $\text{eff}_{c\hat{\beta}}$ is the contrast efficiency and σ^2 is residual variance. Thus doubling efficiency cuts the $Var(c\hat{\beta})$ in half [Liu et al., 2001], increasing statistical power. Efficiency's direct link to $Var(c\hat{\beta})$ is beneficial, but it lacks a benchmark threshold that would allow the evaluation of a single study design and can only rank study designs. While VIF has a benchmark threshold and assesses the model for a single study design, we will establish that it may not reflect $Var(c\hat{\beta})$ as expected.

In a linear model, $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$, the VIF for $\hat{\beta}_j$, the parameter estimate for the jth regressor, is the ratio of $Var(\hat{\beta}_j)$ in a model with all p regressors to a model only including X_j and an intercept. This compares a "worst-case" scenario (maximal collinearity) to a "best-case" scenario (no collinearity) where less than a 5-fold increase in variance is a common benchmark. However, this best/worst definition may not capture correlations between task conditions, where the ideal case would be uncorrelated conditions, and the worst case would reflect true correlations in the data. These definitions only align when between-condition correlations are reflected through between-regressor correlations, achieved by having a single regressor for each condition. An equivalent model with orthogonal regressors will have the optimal VIF value of 1 but fails to capture the impact of condition correlations on contrast variance. Appendix A illustrates this point with a 2 condition task and Appendix B presents the recommended model and introduces a procedure for estimating VIFs for contrasts of parameter estimates, or contrast VIF (cVIF).

fMRI researchers often aim to remove all correlations between regressors, but some collinearity is expected

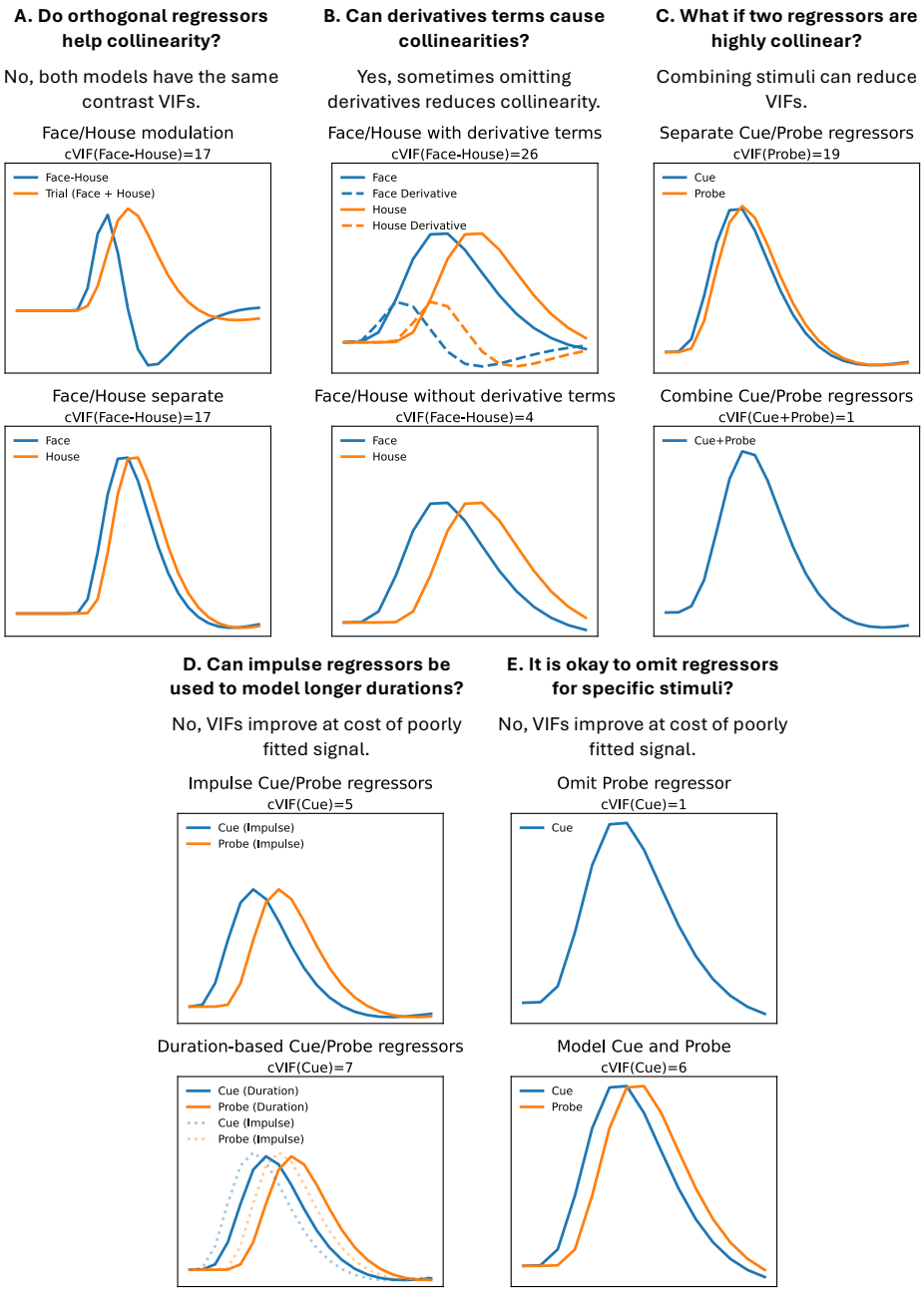

Figure 1: Different modeling strategies used when collinearity is encountered. Each panel represents a different scenario where the top model is a common starting point and the bottom model is an equivalent (panel A) or improved modeling strategy (Panels B-E). Only a single trial is shown for each stimulus for the purposes of illustrating the collinearity, but the cVIFs are based on a 50s long scanning run. The cVIF for the contrast of interest is displayed at the top of each panel and is based on the new algorithm presented in Appendix B.

and acceptable. Figure 1 evaluates common strategies for handling collinearity, from ineffective to acceptable to unacceptable. Each panel poses a question with answers for two competing designs, where the top design is a typical starting point and cVIFs, estimated using the algorithm from Appendix B, are shown in the

regressor figure titles. For a clear illustration of the collinearity for each design, the figure only shows a single presentation of each stimulus, but the cVIFs are based on as many trials that would fit into a 50s long period. Panel A of Figure 1 addresses whether using statistically orthogonal regressors (i.e., correlation between regressors is 0) is beneficial when facing collinearity. As mentioned earlier and in Appendix A, the pair of models in this example are equivalent, with identical inferences for all contrasts and, thus, equal cVIFs for Face-House (cVIF=17). The traditional, regressor-based VIFs of the orthogonal regressor model are 1, regardless of how much the Face and House signals overlap and will not reflect when the collinearity is impacting the properly estimated cVIF for Face-House. While not necessarily harmful, this may imply a false superiority of one model over another and may mask power losses due to poor study design when high task collinearities are present.

Panel B of Figure 1 shows an example where omitting a regressor is acceptable. Temporal derivative terms can introduce high collinearity when added to a model when stimuli are short and temporally close. The purpose of adding a derivative is to allow the fitted time course for a stimulus to shift in time forwards or backwards by a small amount, accounting for misspecifications in the specified canonical HRF shape. If two stimuli are very close then a linear combination of the main regressor and derivative of one stimulus may allow enough of a shift to closely approximate the main regressor of the other stimulus, resulting in a high VIF. If the derivative terms are not used for amplitude correction—which requires a nonlinear combination of the original regressor and its derivative [Calhoun et al., 2004]—they serve no purpose and can be omitted to reduce collinearity. If there is a shift in the true HRF compared to the assumed HRF, omitting the derviative can introduce some degree of mismodeling. However, the benefits of reducing the VIF likely outweigh this risk. If a severe mismodeling of the onset is suspected for a given task and the derivative regressors are omitted, then the reported results should clarify that bias from mismodeled onset signal may be present. Alternatively, the design optimization stage can focus on models that allow for modeling of the derivative and use of the derivative parameter as shown in [Calhoun et al., 2004] or an alternative modeling strategy as shown in Lindquist and Wager [2006] to address this situation. In this example, removing the derivatives reduces the cVIF from 26 to an acceptable 4.

For regressors with higher VIFs (> 10), it is acceptable to omit the corresponding parameters with high VIFs from hypothesis testing while retaining the regressors in the model. Another acceptable solution is to combine multiple collinear stimuli into a single regressor (Panel C, Figure 1). In this example, regressors for a cue and a successive probe are highly collinear, prohibiting the study of cue and probe, independently. Instead, they can be treated as a single cue/probe event (starting at cue onset and stopping at probe offset), which removes the collinearity with an attendant loss in interpretation, since the source of the activation cannot be distinguished between the two components.

A collinearity reduction strategy that should be avoided is modeling a condition with an impulse regressor instead of the actual stimulus duration (Panel D, Figure 1). The figure illustrates how mismodeling the duration has a large impact on the location of the peak of the activation. The true duration of a stimulus should be used to ensure the condition-related signal is fully captured by the regressor to prevent biasing of other parameter estimates, which will be shown in our simulated data results. Of course there may be limitations in knowing the true neural activation duration as well as the potential that it may vary across the brain [Summerfield et al., 2006], but this should be done to the best of one's knowledge about the task. If the true activation duration is thought to be less than the stimulus duration, then it is still likely necessary to model the remaining stimulus duration's activation in a secondary regressor or design the task so the duration of the stimulus aligns with expected duration of the neural signal.

Modeling the duration is related to the importance of modeling response times, since response time variability may mirror neural process duration variability. Unmodeled response times that differ between conditions can lead to inflated error rates in condition comparisons. This is typically remedied by adding a single response time regressor [Mumford et al., 2024].

Omitting regressors from the model (Panel E, Figure 1) is a detrimental practice to remedy collinearity due to its potential to significantly bias estimates. While the impact of unmodeled confounds in linear regression is well established Greene [2003], our simulations further demonstrate that this can be a substantial source of bias. For the MID task, some of these biases were anticipated, which led to unwritten rules about omitting some contrasts from analyses.[1] Unfortunately, there can be inconspicuous biases that are not revealed without a deeper analysis of the timings of the stimuli, as we carry out here. Examples of obvious and unexpected biases will be illustrated in our simulation analyses. More generally, the omission of known variables can violate a basic assumption of linear modeling, known as "exogeneity" in the econometrics literature, which states that the error term in the model should be uncorrelated with any independent variable (whether or not it is included in the model). If this fails, as it can when an omitted variable is correlated with both an included independent variable and the dependent variable, then the statistical estimates can be invalid [Antonakis et al., 2010].

The model we propose has acceptable VIFs with the ABCD implementation of the MID task, often lower than the original model used by the ABCD consortium. Our model corrects all biases and removes the need to limit the contrasts that can be studied in the MID task. Our real data analysis will compare the model results, but care should be taken not to interpret activation increases as "power" if a potential for bias has been established in the simulation analyses, since the concept of power is only interpretable for unbiased

[1] This insight comes from personal conversations and an instance where a reviewer requested that one of the authors remove contrasts in MID task analyses due to potential between-condition bias from an unmodeled fixation stimulus. We were unable to find a specific reference describing this bias.

estimates.

Next, we review the history, design and traditional modeling decisions of the MID task. We do not discuss the mixed selectivity of brain regions for distinct components of the task, which are outside the scope of this paper [Oldham et al., 2018, Bartra et al., 2013].

The history and design of the MID task

The MID task was inspired by preclinical work that suggested dissociable affective responses to "reward anticipation", "reward feedback" and error-related "learning" [Knutson and Wimmer, 2007, Schultz, 1998]. Knutson's original task design focused on mapping these distinct stages of reward-guided behavior to their neural correlates in humans across the distinct temporal events in the task: anticipation and feedback [Knutson et al., 2001]

In the early research, Knutson and colleagues reported that the reward anticipation stage was associated with heightened activity within the ventral striatum with a particular emphasis on the Nucleus Accumbens (NAc), while reward feedback was associated with heightened activity of the ventromedial prefrontal cortex (vmPFC; e.g., [Knutson et al., 2003]). Researchers administer the task due to its postulated ability to distinguish between distinct reward valence systems, positive to *approach* and negative to *avoid* [Knutson et al., 2014]. As a result, the MID has been adopted by many groups as a robust measure of incentive motivation processing ([Bjork, 2020, p. 422]) and included in the ABCD Study based on its ability to delineate processes of anticipation and receipt of reward ([Casey et al., 2018, pp. 46-48]).

The trial-by-trial structure across MID task administrations is comparatively similar to the version described in Knutson et al. [2001] and the ABCD study [Casey et al., 2018]. Participants are instructed that on each trial they have the opportunity to *win* or *avoid losing* a large or small amount of money, or no money is at stake. A cue is first presented (henceforth, Cue), which notifies the participant of the trial type - Win, Don't Lose or Neutral (i.e., no money at stake for that trial), as well as the magnitude of the potential gain or loss (e.g., \$5/\$0.20 in the ABCD study). After the presentation of this Cue for a fixed amount of time (2000 ms in the ABCD study), a fixation point (henceforth, Fixation) then appears for a fixed [Yau et al., 2012] or variable amount of time (from 1500 to 4000 ms in the ABCD study). After the Fixation, a probe appears (henceforth, Probe) for a performance contingent duration (ranging from 150 to 500 ms in the ABCD study); participants are instructed to respond to the Probe via button press as quickly as possible. A successful Probe response occurs if the participant responds during the Probe presentation window; this results in successfully winning or avoiding a loss. The duration of the Probe presentation is adaptively modified to achieve target accuracy. In the ABCD study, the probe window is adjusted after

every third trial based on the running Probe accuracy rate of the last six trials across all Cue conditions with a targeted accuracy of 60% [Chaarani et al., 2021, Casey et al., 2018]. In other instantiations of the MID task the duration may be adjusted within trials of each cue type [Wu et al., 2014]. Following the Probe, the participant receives feedback (henceforth, Feedback) on whether the participant was successful or unsuccessful (too fast/too slow) in responding during the allotted Probe window. The Feedback stimulus indicates the resulting reward for that trial using a fixed [Knutson et al., 2003] or variable duration [Casey et al., 2018]; in the ABCD study, the Feedback duration is 1950 ms minus that current trial's Probe duration.

Variability in fMRI time series models of the MID task

Each trial during the MID task aims to temporally separate distinct components of reward processing [Knutson and Wimmer, 2007]. The proposed theoretical framework is that the anticipatory phase serves as the prediction for the anticipated reward (or expected value) and the outcome serves as the Feedback or prediction error.

While the Cue and/or Fixation components of the task are commonly modeled as "anticipatory", the definition of this construct remains variable in the MID task literature [Demidenko et al., 2024]. Whereas the Cue specifies the reward type and amount for a given trial, the Fixation is a preparatory signal for the behavioral response during the Probe. As a result, some have referred to the Cue and Fixation as the anticipation of prospect, or the *motoric anticipation* [Andrews et al., 2011, Balodis et al., 2012, Garrison et al., 2017], separating it from the *anticipation of incentive* which occurs during the phase after the Probe and before the Feedback in versions of the task where a jitter is included between the two components. If the Cue serves as the expected incentive of a given trial and the Fixation serves as a preparatory response for that value, these adjacent components may contain overlapping but distinct cognitive processes. As a result, there is variability in how time series models of the MID task are specified, with differing definitions of anticipation and different stimuli being left unmodeled.

Typically, the time series model of the MID task only includes regressors for the anticipation and Feedback components, while leaving other stimuli unmodeled [Büchel et al., 2017, Yau et al., 2012, Srirangarajan et al., 2021]. The Probe is frequently omitted from the time series model, with some exceptions [Beck et al., 2009, Bjork et al., 2011, Beck et al., 2009, Knutson et al., 2005]. Anticipation has been defined variably, including: 1) start and duration of the trial's Fixation stimuli, leaving Cue unmodeled [Knutson et al., 2003]; 2) the start and duration of the trial's Cue stimuli, leaving Fixation unmodeled [Joseph et al., 2016]; 3) an impulse response at the start of the trial's Cue stimuli, leaving Fixation unmodeled [Hagler et al., 2019]; or 4) the start of the trial's Cue stimuli and *combined* duration of the trial's Cue and Fixation stimuli [Balodis et al., 2012,

Martz et al., 2016]. In some cases, like in the ABCD study [Hagler et al., 2019], researchers have included temporal derivatives for each of the modeled stimuli [Joseph et al., 2016]. It is rare for response times to be modeled in the MID task, although RT adjustments to the anticipatory signal were made in Knutson et al. [2005]. This variability in definitions and model specifications makes it challenging to compare the model derived estimates, which can vary substantially across the different specifications [Demidenko et al., 2024].

Another variable modeling strategy involves the definition of Miss trials, which is rarely documented in manuscripts, so we have relied on personal communications with other MID researchers for clarification. Hit trials are consistently defined as a button press during the Probe phase, but the other two outcomes—Too Soon (a button press before the Probe phase) and Too Slow (a button press absent or after the Probe offset)—are used inconsistently. Some researchers classify both Too Soon and Too Slow as Miss trials, while others define Miss trials as Too Slow only, discarding Too Soon trials or modeling them as a separate regressor of no interest [Büchel et al., 2017]. Leaving these unmodeled has the potential for them to bias estimates of the subsequent feedback component.

Methods

A subset of the ABCD dataset was used for both simulated and real data analyses. Task timings and participants' behavioral responses were used to generate design matrices and simulate fMRI data, which were analyzed using two modeling strategies. These models were then applied to the real data, allowing for a comparison of group activation maps from commonly used MID contrasts.

ABCD Data

Participants The ABCD sample used here is from the raw ABCD-BIDS Community Collection (ABCC: [Feczko et al., 2021]), which is accessed via the Minnesota Super-computing Institute and Amazon Web Services. The ABCD study protocol acquired structural and functional volumes across 21-sites and three scanners [Casey et al., 2018]. The focus here is on the 2-year follow-up data for two reasons. First, earlier versions of the MID E-prime procedure administered during baseline did not collect reaction times (RTs) for miss trials. Second, the sample used here is a sub-sample assessed for quality control in a previous analysis [Demidenko et al., 2024] such as successful field map correction, co-registration between anatomical and functional data, mean framewise displacement $< .9$, and availability of e-prime data and adequate task performance. For the final analytic sample (N = 500), the average age of the participants is 12.03 years-old (SD: 0.66) and 47% are Female (N = 235). All participants' data are from SIEMENS scanners across 13 different ABCD sites.

Monetary Incentive Delay task These analyses focus on the modified version of the MID task [Knutson et al., 2001] that is administered in the ABCD study protocol [Casey et al., 2018]. A trial of the MID task includes Cue, Fixation, Probe and Feedback components. There are five Cue types: Win $5 (Large Win), Win $0.20 (Small Win), No Money at Stake (Neutral), Don't Lose $5 (Large Loss), and Don't Lose $0.20 (Small Loss). Based on the subject responses, Feedback is classified as either a Hit (successful probe window response) or Miss (Too Soon/Too Slow response).

During the development of these analyses we identified issues with the ABCD acquisition and timing files. First, since the *.Duration* columns in the E-prime files are the requested and not the true stimulus screen-display durations, we use the *.OnsetToOnsetTime* column as the true duration [Psychology Software Tools, 2024]. In the case of the Feedback stimuli there isn't a *.OnsetToOnsetTime* column so we use the difference between the onset of the next trial's Cue and the onset of the current trial's Feedback as the duration, which is consistent with the *.OnsetToOnsetTime* definition [Psychology Software Tools, 2024]. Second, the original response time estimates within the Eprime files, denoted "OverallRT" were based on flawed duration values for Miss trials, so we recomputed them using updated trial durations. This adjustment affected only the Too Soon and Too Slow response times, when present. A trial is classified as Too Soon if any button press occurs before the Probe. While it may seem counterintuitive to report a response time for such trials, this depends on the timing and number of button presses. If only a single button press occurs, no response time is recorded. However, if additional button presses occur after the Probe onset, the response time of the first press is recorded for that trial. Lastly, as an aside, trial onset timing values for a subset of ABCD data were misspecified due to an incorrect estimation of calibration volumes in all GE data; see Supplemental Section S1 for a more detailed description of this issue. This issue impacts all ABCD and ABCC GE data released prior to May 2023 and is expected to be fixed in the upcoming ABCD Release 6.0.

The time series models studied are defined in terms of either impulse functions (i.e. Dirac delta) or boxcar functions convolved with a double gamma hemodynamic response function (HRF), specifically the `spm_hrf` function from Nilearn [Abraham et al., 2014]. In the ABCD study a Miss is defined as either Too Slow or Too Soon. Here, we compare two models:

- **ABCD model** [Hagler et al., 2019]. All regressors are based on impulse functions at stimulus onsets. There are a total of 5 Cue regressors (large loss, small loss, large win, small win and neutral) and 10 Feedback regressors (each cue type split by hit/miss). Each of these 15 regressors is paired with its temporal derivative for a total of 30 task-related regressors (left panel, Figure S2).

- **Saturated model**. In our proposed model, regressors are based on boxcar functions using the trial onsets and durations. There are 5 Cue regressors split by cue type (duration: 1781-2039 ms), 5

Fixation regressors split by cue type (1500-3666 ms), 1 Probe regressor (172-497 ms), 1 response time (RT) regressor (uses Probe onset and the RT as the duration, between 105-1581 ms) and 10 Feedback regressors split by cue type and hit/miss (1473-1789 ms) (right panel, Figure S2). Note, a single Probe regressor is used since splitting Probe according to Cue type resulted in the Feedback regressors having VIFs that exceed 10.

An illustration of the two models and the estimated components are presented in Figure 2. Panel A differentiates the modeled components for the ABCD (dashed) and Saturated models (solid line). Note, while not shown in the figure, each flagged* regressor in the ABCD model also includes its temporal derivative. Panel B specifies how each onset and duration of the task stimuli are modeled at the subject-level design matrices. The width of the rectangle indicates the duration for each associated component that is modeled. As an example, Panel C provides whole brain maps that can be estimated for each of the model types.

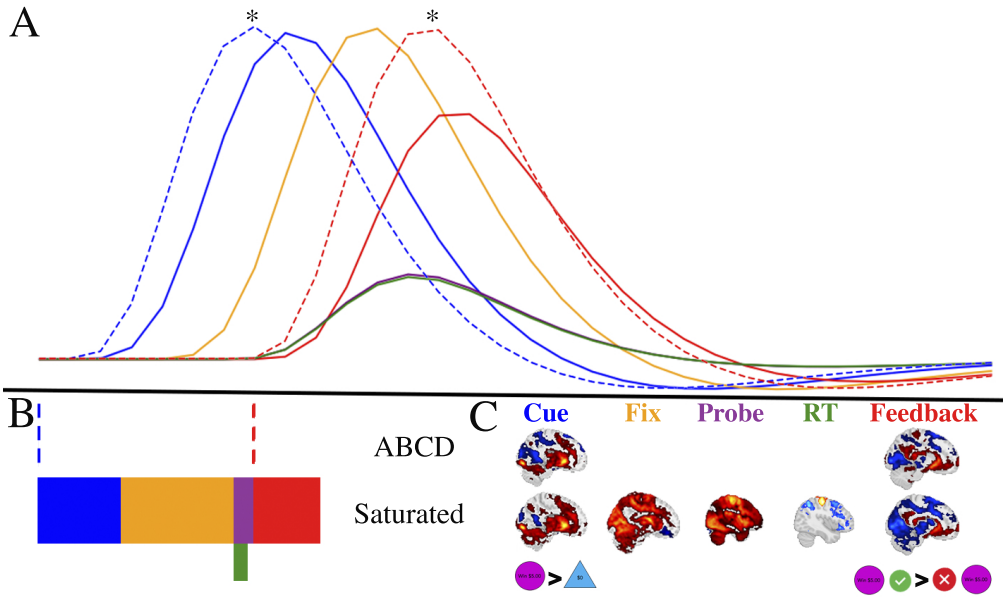

Figure 2: Illustration of the (A) estimated hemodynamic response function for task components, (B) the onsets and durations and the (C) estimated maps between the ABCD (dashed) and Saturated models (solid line). The example activation maps are subset from Figures 5, S12 and S11. Regressors with * in Panel A also include a temporal derivative that is not shown (see Panel B in Figure 1 and S2).

Simulated data analysis

Code for all simulations is available at `github.com/poldrack/MID_simulations`. Design matrices were constructed by concatenating each subject's two runs into a single run. Synthetic data were generated by multiplying the Saturated model design matrix by the desired true signal components, and then adding independent Gaussian noise to the true signal. The time resolution matched that of the real data (800 ms).

Independent noise is not a plausible assumption for fMRI data, but it does not impact the overall simulation results and allows for the use of a simple ordinary least squares model here. Between-participant variability was introduced by sampling the true signal components from a Gaussian distribution with the intended mean; the standard deviation of the between-participant and within-participant variability were set as equal to 1. For each simulated data set, aside from the null data simulation, only a subset of parameters had nonzero values as shown in Table 1. Parameter values were chosen such that the power of detection, for that parameter, was approximately 80% for the group average. The contrasts we studied are shown in Table 2. For each parameter setting, a total of 1000 data sets containing 500 subjects each was simulated and analyzed. Power was estimated for contrasts involving nonzero parameters while error rate and bias were estimated for contrasts that were null for that simulation setting.

Condition(s) of focus	**Nonzero parameters**
Null model	(all betas 0)
Cue (Win)	Cue LargeWin = 0.15, Cue SmallWin = 0.15
Fixation (Win)	Fixation LargeWin = 0.15, Fixation SmallWin = 0.15
Probe effect	Probe = 0.55
Response Time	Probe: RT = 0.3
Feedback	Feedback LargeWinHit = 0.18, Feedback LargeWinMiss = 0.18

Table 1: Simulation settings. Each row corresponds to a different simulated data set, where the nonzero parameters are specified in the second column.

Contrast Name	**Contrast Definition**
Cue: W-Base	.5 × Cue LargeWin + .5 × Cue SmallWin
Cue: W-Neut	.5 × Cue LargeWin +.5 × Cue SmallWin -1 × Cue Neutral
Cue: LW-Neut	1 × Cue LargeWin -1 × Cue Neutral
FB: WHit-NeutHit	.5 × Feedback LargeWinHit +.5 × FeedbackSmallWinHit -1 × FeedbackNeutralHit
FB: LWHit-LWMiss	1 × Feedback LargeWinHit -1 × Feedback LargeWinMiss

Table 2: Contrasts of interest in the simulated data analyses.

Real data analysis

Preprocessing The data were preprocessed using fMRIPrep v23.1.4 [Esteban et al., 2022]. In short, a T1-weighted image was corrected for intensity normality, skull-striped via Nipype's antsBrainExtraction.sh workflow and spatially normalized to MNI152NLin2009cAsym standard space using non linear registration via AntsRegistration. For each of the two MID BOLD functional runs, the following sequence of steps were performed. A reference volume and its skull-stripped version were generated, a fieldmap was estimated and aligned with rigid-registration to the target EPI (echo-planar imaging) reference run and each BOLD run was corrected for signal distortion using top-up. The BOLD reference image was co-registered to the T1w image using bbregister using boundary-based registration [Greve and Fischl, 2009] with six degrees of freedom. The BOLD time-series were resampled into standard space, resulting in a preprocessed BOLD run in MNI152Nlin2009cAsym space. Head-motion parameters (transformation matrices and six corresponding rotation and translation parameters) were estimated before any spatiotemporal filtering using mcflirt [Jenkinson et al., 2002] and expanded with the inclusion of temporal derivatives and quadratic terms for each [Satterthwaite et al., 2013]. The time-series data were filtered using a discrete cosine filter with 128s highpass cut-off. For each BOLD run, an associated confound file was generated with the estimated motion parameters and the cosine basis set that were used in subsequent subject-specific GLMs.

Within-Run and Within-Subject Models Within-run and within-subject models were estimated using *Nilearn* 0.9.2 [Abraham et al., 2014] in Python 3.9.7. Group-level models and inference were estimated using FSL's *randomise* (1000 permutations) with Threshold-free Cluster Enhancement (TFCE).

For each of the two runs, the within-run model fits a GLM to the subject-specific time series data using the function `FirstLevelModel` in *Nilearn*. In addition to the task regressors described above for the two models, cosine regressors calculated by *fMRIPrep*, corresponding to a 128s highpass filter cut-off, and the x, y, z translations, rotations and their derivatives (12 motion parameters), were included in the subject-specific design matrix. The time series were pre-whitened using an AR(1) model and spatially smoothed using a 5 mm FWHM smoothing kernel. For each contrast of interest, *Nilearn*'s `compute_contrast` was used to obtain a contrast estimate (`effect_size`) and its estimated variance (`effect_variance`). The within-subject model estimates average contrast estimates across the two runs. For within-subject estimates we used *Nilearn*'s `glm.compute_fixed_effects` with the precision-weighted option to obtain a weighted-average of the contrast estimates across runs.

For each contrast, a 1-sample t-test of the within-subject contrast estimates was estimated using FSL's *randomise*. Inference for a two-sided t-test of the mean was used to threshold statistical maps. For display

purposes the t-statistics were converted to Cohen's d estimates using:

$$d = \frac{t}{\sqrt{n}}.$$

Cohen's d values are displayed for voxels that were found to have $p < 0.05$ for a two-sided t-test of the mean, using the TFCE-based p-value.

Region of Interest (ROI) Estimates ROI and model-specific contrast estimates were extracted for the Left and Right NAc. The Left and Right NAc are from the thresholded Harvard-Oxford subcortical atlas (*nilearn* Harvard-Oxford: sub-maxprob-thr50-2mm) using the procedure described in supplemental section S6. To extract ROI estimates from the statistical maps, for each contrast the voxels were masked and averaged within the Left and Right NAc.

Results

Simulation Study

Figure S3 displays the cVIFs for regressors and contrasts using our new cVIF algorithm for contrasts (Appendix B). Unexpectedly, even without an ITI in the ABCD version of the task, the Saturated model with regressors for all MID task conditions does not exhibit problematic collinearity. Most cVIFs are near or below 5, but one contrast, CUE: W-Base, is near 7 for the Saturated model. While elevated VIFs can suggest reduced power, inferences remain valid. When VIFs exceed the commonly accepted threshold of 5 for a contrast of interest, we recommend the researcher carefully inspect the data for outliers and the outliers should be omitted from further analyses. In our experience contrasts with VIFs exceeding 20 are too prone to outliers and we do not include them in analyses. We did not include derivative terms in our Saturated model as their addition led to 30 of the 49 model parameters having VIFs larger than 20.

Efficiency differences between the two models are modest across regressors and contrasts (Figure S4). Ratios of Saturated to ABCD model efficiencies show that ABCD variances range from 0.74 to 1.26 of the Saturated model variances. Given that between-session and between-subject variances typically dominate overall variance in group-level models, these small efficiency differences are unlikely to affect group-level results. The biases, however, described below, present a larger issue.

Recall that the ABCD model excludes regressors for Fixation, Probe, and RTs, leaving unmodeled variability that can bias other model parameters. Additionally, if these unmodeled stimuli vary in duration across trial and outcome types, this can unevenly bias parameters, as duration differences in the neuronal signal may appear as amplitude differences in the BOLD signal [Mumford et al., 2024]. Figure 3 shows

that Probe, Fixation, and RT durations differ between trial outcomes (Hit/Miss), indicating a potential for biases on contrasts between Hit and Miss. To clarify details about the RTs, the measured RTs are button presses after Probe onset, with Too Soon RTs indicating a secondary press within a trial and Too Slow RTs corresponding to presses after Probe offset. On average, Probe durations are 26-33ms longer for Hit trials, Fixation durations are 63-160ms longer for Hit trials and RTs are 78–100ms longer for Miss trials.

Figure S5 further breaks down Miss trials into Too Soon and Too Slow, highlighting notable differences: Compared to Too Slow trials, Fixation durations are much longer (700ms) for Too Soon trials, Probe durations are slightly longer (10ms) for Too Soon trials and RTs are shorter (42ms) for Too Soon trials. On average, trials are 57% Correct, 36% Too Slow, and 7% Too Soon. Given these differences as well as differences in the cognitive processes that are occurring during Too Slow and Too Soon, a more specific model would separate Feedback regressors into Correct, Too Soon, and Too Slow, with a focus on comparisons between Correct and Too Slow while Feedback for Too Soon would be treated as a nuisance regressor. We have not explored such a model, for the sake of brevity, but we suggest that it be considered in future MID research.

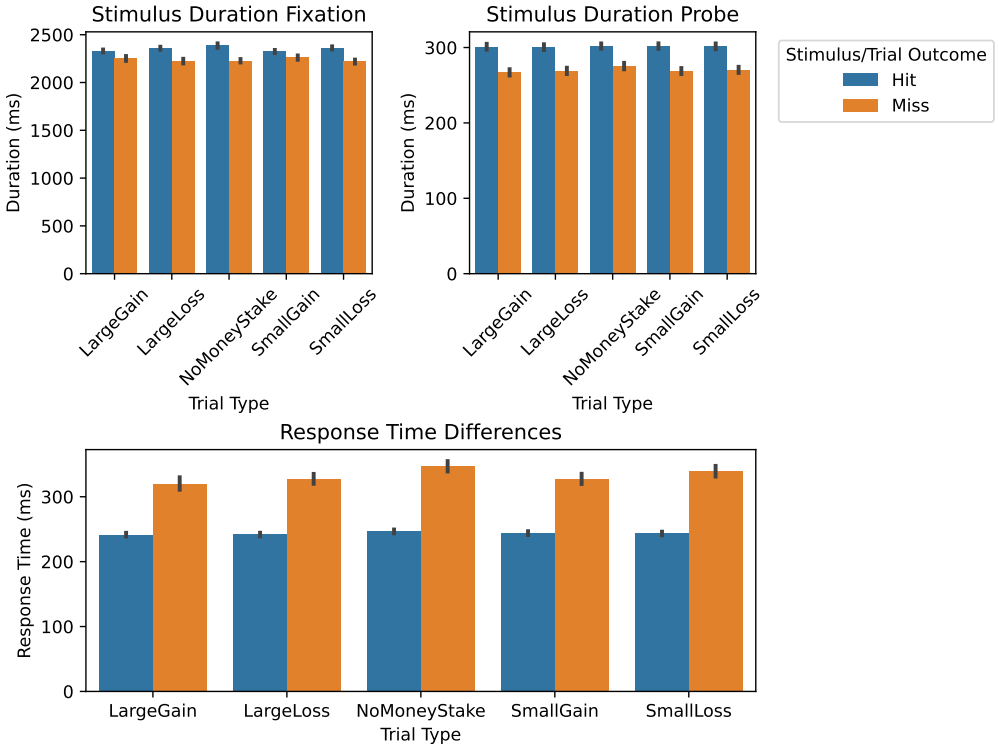

Figure 3: Stimulus durations and response times split by trial type and trial outcome. Differences in duration or RTs between trial outcomes indicate the potential for contrasts comparing hits and misses to be biased.

Figure 4 displays the average group Cohen's D values for the contrasts of interest across simulation settings for both models. In all cells where values are displayed, the contrast estimate should be 0, so

nonzero values reflect bias. A dark gray outline surrounding the cell indicates the Type I error was larger than 0.066, which is the smallest proportion whose 95% confidence interval does not contain 0.05, the desired error rate. The error rates and biases for all parameters and contrasts are shown in Supplemental Figures S6 and S7, respectively. In this case elevated error rates indicate that a sample size of 500 subjects was large enough to detect biases present in the parameter estimates. In some cases biases are present without inflated error rates, but at larger sample sizes the biases can impact error rates. For example, a bias here of .5 would have 10% more power to be detected in a sample size of 2000 compared to 500.

Figure 4 shows controlled biases for the Null setting for both models and the Saturated model has controlled biases for all settings. When signal is introduced for the Cue stimuli during Win trials (second row, Figure 4), bias is present for FB:WHit-NeutHit, indicating the impulse regressor's inability to capture all of the cue-based activity in the ABCD model. As expected, since there are no Fixation regressors in the ABCD model (third row, Figure 4), when Fixation signal is introduced for Win trials the unmodeled signal biases the CUE:LW-Neut, Cue:W-Neut and FB:WHit-NeutHit. Importantly, there is negative bias present for Cue:W-base, due to Fixation signal contributing to the Baseline estimate but it does not impact error rates at this sample size. The Probe duration differences introduce a positive bias in the ABCD model for FB:LWHit-LWMiss (4th row, Figure 4). This is due to the Probe duration differences between Hit/Miss trials. Adding RT signal introduces a small amount of bias that does not impact error rates at this sample size. When Feedback signal is added, since the impulse regressor of the ABCD model does not capture all of the signal, the remaining signal introduces small biases of the cue-based estimates.

Real-world Data Analyses: Effect of model choice with the MID task

The within-run, within-subject and group-level analytic code, and the Jupyter notebook used to generate the associated tables and figures from the estimated maps are available at https://github.com/demidenm/mid_rt_bold. All within-run, within-subject and group maps use data from N = 500 ABCD sample participants. The resulting randomise t-stat, f-stat and Cohen's d estimated images across each contrast for the group and paired t-test between models are available on Neurovault: https://identifiers.org/neurovault.collection:18697.

Behavioral Results Participant's Probe response accuracy across Cue conditions and mean RTs across feedback conditions are reported in supplemental Figure S11 and in 3, respectively. As expected, probe response accuracy differs across Cue stimuli and mean RT values differ across Hit/Miss Feedback stimuli.

Comparing Models: Whole-Brain Activation Results The contrasts of focus for these real data analysis are defined in Table 3, while eighteen additional contrasts and their results are included in the

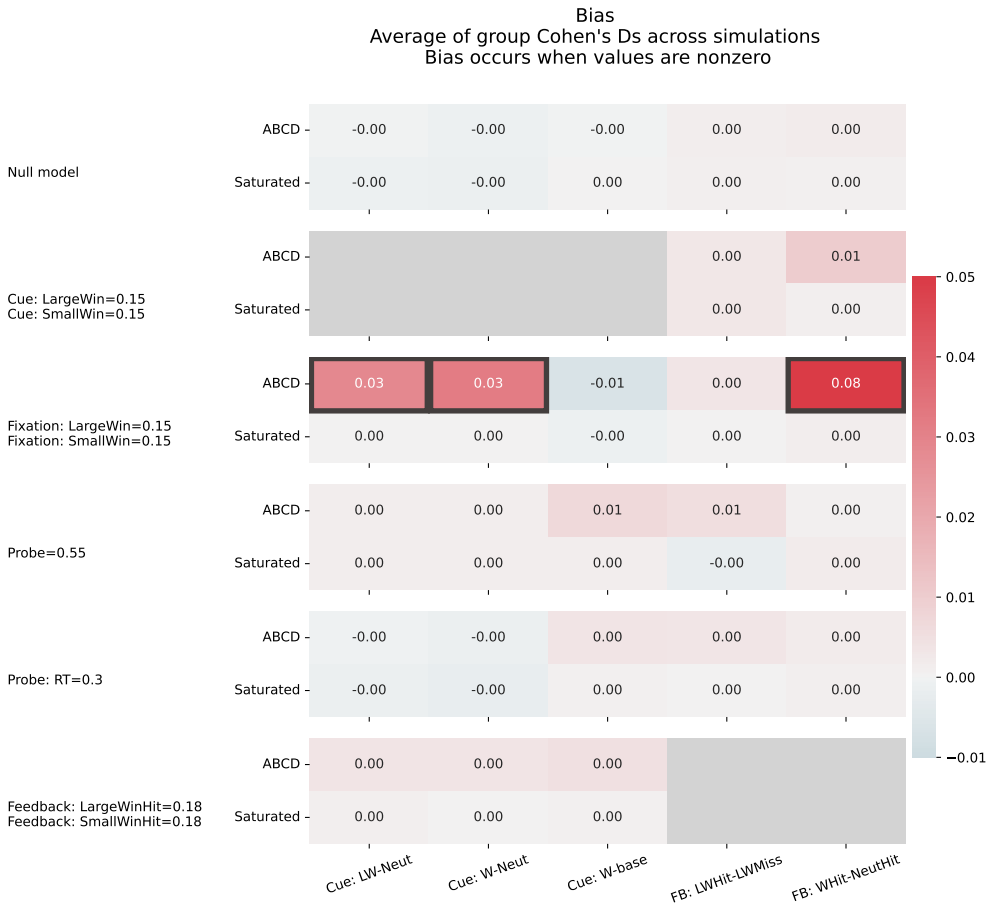

Figure 4: Bias for contrast estimates across simulation settings and models. Gray squares around cells indicate the Type I error rate was inflated, while the color reflects the average t-statistic, where a value of 0 reflects no biasing. Each row corresponds to a different model where the contrast estimate bias is shown for the ABCD (top) and Saturated (bottom) models. The solid gray cells indicate that a parameter involved in that contrast was nonzero and so error rate is not estimated.

Supplemental materials S8.

Across the Cue and Feedback contrasts for the ABCD and Saturated models, the thresholded group average activation maps are presented in the first two columns of 5 (Voxelwise Cohen's d estimates displayed in voxels with TFCE $p < 0.05$ for a 2-sided t-test; 1000 permutations). Consistent with the literature, the Cue contrasts (rows 1 - 3 in Figure 5) demonstrate a broad activation of striatal, insula, visual and motor areas. Likewise, consistent with the literature, the within condition Hit versus Miss contrasts (rows 4 & 5 in Figure 5) demonstrate activation in mPFC and striatal areas and deactivation in the insular, visual and motor areas. The between condition Hit contrast (row 6) demonstrates a broad deactivation in striatal, insular, visual, insular and motor areas.

The ABCD and Saturated models meaningfully differed in the estimated BOLD activity at the group-level across the seven contrast types (The complete list of contrasts are presented in supplemental Figure

Contrast Name	Contrast Definition
Cue: LW-Neut*	1 × Cue LargeWin - 1 × Cue Neutral
Cue: LL-Neut*	1 × Cue LargeLoss - 1 × Cue Neutral
Cue: LW-Base	1 × Cue LargeWin
FB: LWHit-LWMiss*	1 × Feedback LargeWinHit - 1 × FeedbackLargeWinMiss
FB: LLHit-LLMiss*	1 × Feedback LargeLossHit - 1 × Feedback LargeLossMiss
FB: LWHit-NeutHit	1 × Feedback LargeWinHit - 1 × FeedbackNeutralHit
FB: LWHit-Base	1 × Feedback LargeWinHit

Table 3: Primary contrasts of interest in real data analyses.
* Contrasts included with ABCD data releases Chaarani et al., 2021.

S10). As is illustrated in the paired t-stat difference map in the right column in Figure 5, the ABCD and the Saturated model meaningfully differ in activation magnitude in the visual, motor, insular and striatal regions across the contrast maps. These differences reached small to medium effect sizes in multiple regions. For example, the mean effect size (Cohen's d) for the activation reduction within the NAc region of interest during Feedback ranged from d = -.68 to -.37 and d = -.63 to .-40 for Left and Right, respectively, across four of the within-condition Hit versus Miss contrasts, and d = -.01 to .01 and d = -.02 to -.01 for Left and Right, respectively, across Large Win Hit > Neutral Hit and Large Win Hit > Implicit Baseline contrasts.

In supplemental Figure S12, we present the Cue and Fixation contrast maps for the Saturated model. For the Fixation component, in the between-condition contrasts (e.g., Large Win>Neutral), there is heightened activity in the visual, motor, dorsal striatal and posterior brain regions compared to the contrasts for the Cue component. Furthermore, as is illustrated for the Saturated model in Figure S12, the Large Win>Implicit Baseline contrast map is largely negative for Cue and positive for the Fixation component. This illustrates that the Cue and Fixation components differentially impact the estimate of Baseline activity and should be accounted for in subject-specific models.

Supplemental Figure S11 shows activation maps for the average Probe and RT effects. The Probe is associated with very high positive BOLD activity across the motor system and other regions. The Probe RT is associated with positive BOLD activity in the motor cortex and decreased BOLD activity in a broad set of regions including frontal and parietal regions. While present, the magnitude of the RT effect in this ABCD sample is relatively small.

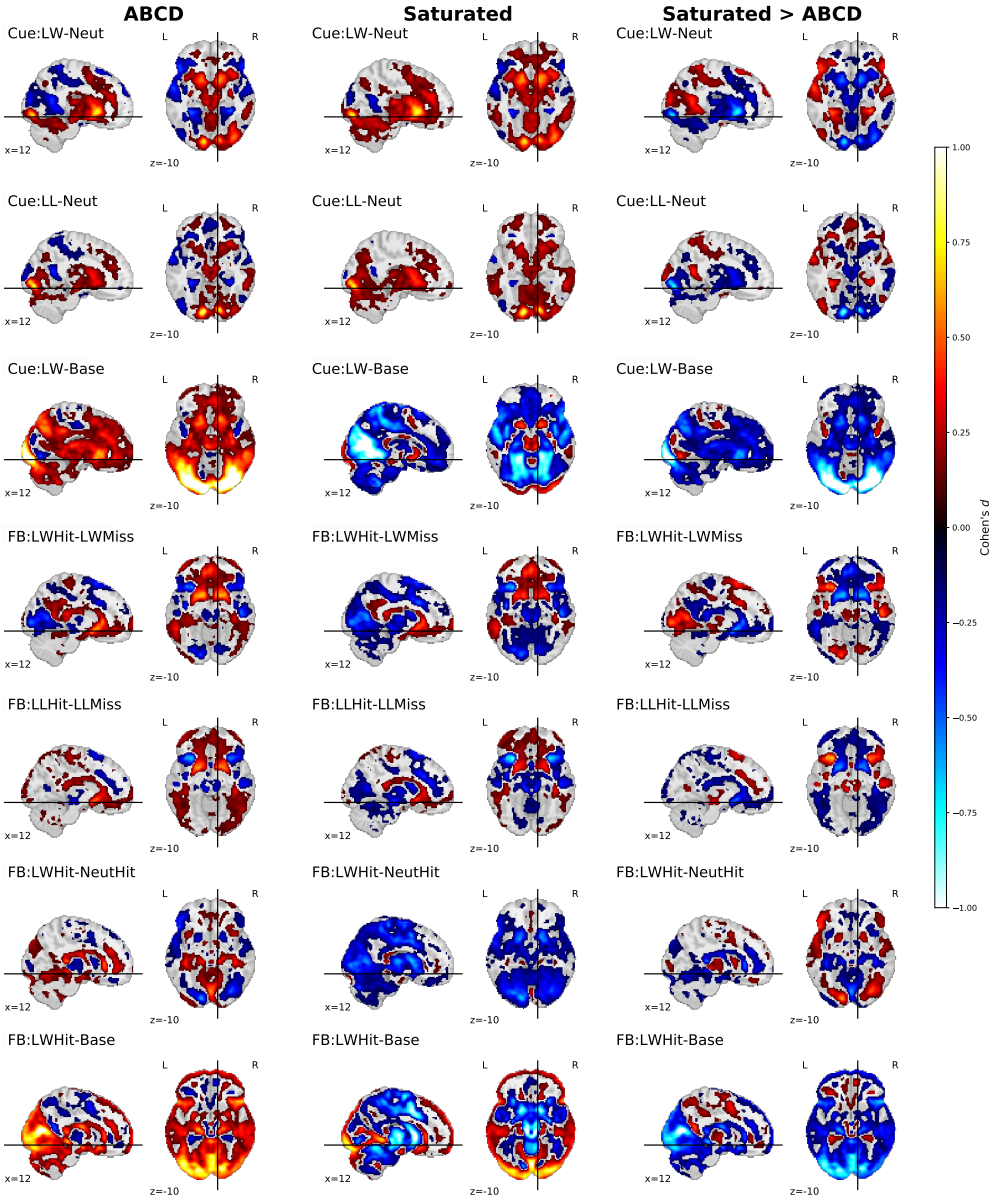

Figure 5: Cohen's d group-level activation maps (voxels selected using TFCE $p < .05$, with 1000 permutations) across a subset Cue (3/5) and Feedback (4/6) contrasts for the ABCD (shown on the left) and Saturated model (shown in the middle) and the Paired statistical difference (Saturated>ABCD; shown on the Right)

Discussion

Given the weak signal in task fMRI, relative to the noise, researchers optimize task designs and subject-level models of the BOLD time series data to reduce collinearity between task regressors and increase power to detect small effects. Although the omission of task components or mismodeling the estimated duration of BOLD activity are common practices in task fMRI modeling, as they may reduce collinearity, they may also

introduce bias into parameter estimates. Using ABCD's MID task, we compared the ABCD and Saturated model using simulations. These simulations affirm previously known biases and identified new biases of contrast estimates in MID task modeling. The Saturated model, which accounts for all task components, their true durations and response times, reduced biases from affecting contrast estimates. Even though there are multiple, adjacent components within a trial, with no resting fixation in between trials, the collinearity of the Saturated model regressors is within the acceptable range. In real-world data, the Saturated model elicits BOLD activity in expected regions for typical Cue and Feedback contrasts, accounting for adjacent task components. After working collaboratively with the ABCD task consortium to study and develop this approach, we propose the Saturated model procedure be used in the modeling of subject-level time series in future ABCD contrast estimate data releases and encourage task fMRI researchers to reassess their modeling decisions that are solely based on collinearity issues.

To date, some contrast estimates of the MID task were expected to incur bias and were thus avoided. For example, MID researchers are often dissuaded from using between-condition Feedback comparisons, such as FB:WHit-NeutHit, due to the potential for the unmodeled stimuli to differ in activation between conditions. In the context of the FB:WHit-NeutHit contrast, this results in a broad deactivation across the brain as a result of difference in activity during the preceding Cue and Fixation component. However, less known is the fact that if unmodeled stimuli vary in duration across trial and outcome types the differences in the neuronal signal may appear as amplitude differences in the BOLD signal [Mumford et al., 2024]. Hence, if the unmodeled Fixation and Probe components differ in durations across Hit and Miss trial, as they do in the ABCD data, this can introduce bias into the Feedback estimates. Simulations confirm that this is the case in the MID task. In the MID task design, a bias arises due to the differing durations of Probe events between Hit and Miss trials and omission of RT regressors biases the FB:LWHit-LWMiss contrast estimate. This illustrates how some subject-level models of time series data, such as in the ABCD model, may produce biased contrast estimates.

Without precise knowledge of the neural signal's onset and offset timings, statistical models inevitably contain misspecifications, leading to some degree of bias. While some MID task experts may view the bias in Cue-based contrasts as less problematic due to its reflection of anticipatory neural signals, biased estimates of Feedback-related contrasts due to influences of unmodeled Fixation, Probe, and Response times are not reflecting the same neural processes as Feedback and thus impact the validity of the contrast. The present work aims to minimize identifiable biases, and the Saturated model achieves this without restricting the range of contrasts typically studied. To address these challenges, we offer recommendations for the broader fMRI research community, particularly for analysts who may lack expertise in fMRI, the task, or neural signal interpretation. Large-scale studies like ABCD, which make activation estimates publicly available, are

invaluable for advancing the field. However, the growing diversity of researchers using these data highlights the importance of providing contrast estimates with minimal ambiguity as those scientists may lack sufficient fMRI and task-specific training to make these distinctions about degrees of bias, instead assuming that the contrasts are truly valid measures of the labeled constructs.

In the real-world data, it is difficult to directly compare the ABCD model and proposed Saturated model. In part, a variety of changes are impacting the statistics including: the previously mentioned biases, parameter estimate interpretation and power differences. The most evident change in parameter interpretation appears in contrasts against baseline. In the ABCD model, the "baseline" activation consists of activation related to Fixation, Probe and RTs and does not reflect a resting baseline (e.g., time at the beginning or end of the run when they are not performing a task). In the Saturated model, the baseline is the actual baseline and so the contrasts comparing a parameter to baseline have very different meanings between the two models. Although contrasts versus baseline in MID studies are rare, they do occur [Andrews et al., 2011, Patel et al., 2013] and are recommended by some [Balodis and Potenza, 2015]. Compounding this problem across MID studies is inconsistency in how anticipation is defined and what unmodeled stimuli are contributing to baseline. Sometimes the Anticipatory signal is defined as occurring during the Cue (leaving Fixation unmodeled) while other times it is defined as occurring during the Fixation (leaving the Cue unmodeled). These two modeling styles will produce estimated anticipation versus baseline effects that are the negatives of each other [Demidenko et al., 2024]. By using the Saturated model, the interpretation of all contrasts in the model will change because they are now adjusted for Fixation, Probe and RT effects. The impact of this adjustment on the resulting statistic values will vary from no impact to increasing or decreasing a contrast estimate while the statistic interpretation is improved and unambiguous in the Saturated model.

Power differences between the two models cannot be assessed, as the parameter estimates differ in meaning between tasks and, more importantly, biased estimates cannot provide meaningful power assessments. Although there are slight efficiency differences between the ABCD and Saturated model, within 25% of each other, these differences are unlikely to affect the group results presented here as meaningfully as the potential biases in contrast estimates and the variation in parameter meanings. Additionally, since between-subject variance is typically a larger contributor to the overall variance than the within-run variance, which is reflected in the efficiency, power is more heavily impacted by sample size than within-run variance reduction [Mumford and Nichols, 2008]. It is likely that a sample size of 50 let alone the large sample size of the ABCD study is large enough to mitigate concerns about small efficiency differences when using the Saturated model while providing estimates with improved interpretability. Typically, smaller samples (20 or fewer) do not exhibit a significant improvement in power when within-subject variance is reduced, as shown in Turner et al. [2019]. However, exceptions may occur, as demonstrated by Nee [2019].

Across the two models, the real-data contrast estimates elicit activity in similar brain regions during the Cue and Feedback components but they meaningfully differ in magnitude. As expected, the ABCD model elicits activity in the direction and magnitude that is consistent with previous contrast estimates by the ABCD task consortium [Chaarani et al., 2021]. The magnitude of the effects in both the ABCD and Saturated models is weaker than when using a combined duration of Cue and Fixation component as illustrated in Panel C in Figure 1 and group-level activation maps in [Demidenko et al., 2024]. However, the Saturated model, by modeling all of the task components, is capable of producing contrasts estimates for each component of the task that are not biased by adjacent components. While the activity in the task-relevant NAc is reduced in the Saturated model, these BOLD estimates are unbiased by preceding stimuli and allows unique attribution of activity to Cue, Fixation, Probe or Feedback components. This offers the opportunity to study more nuanced differences between individuals across the different phases of the task.

It is common for a researcher to repeatedly use a task in multiple studies and some groups may have study designs that were previously found to be optimal in the efficiency analyses carried out when designing the historical studies. We recommend that moving forward, even if it is a task that a researcher has previously designed and worked with, unless the Saturated modeling approach was used, new efficiency analyses with the Saturated model should be carried out. We also recommend using our newly introduced cVIF calculation (Appendix B) to check for collinearity since it is possible that the range of designs studied in an efficiency analysis all have high collinearity and hence the most efficient design of that set will still have detrimental collinearity present.

If data have already been collected and the Saturated model for that task has high collinearity, it is often the case that stimuli that are short in duration and adjacent will need to be combined as shown in panel C of Figure 1. This is often the case for very short cues followed by very short response probes. This does limit the interpretation of the effect size estimated, but the only other alternative is to model the stimuli without performing inferences on their activation estimates. This is a common problem encountered when first translating a behavioral study for use in fMRI as behavioral studies do not have the same timing limitations that are imposed in fMRI data due to the sluggish BOLD response that is impacted by hemodynamic delay.

It has been brought to our attention that time course plot analyses of the MID task suggest that neuronal activation may not remain consistent throughout the entire stimulus duration, particularly for both the Cue and Fixation periods. As mentioned in the introduction, our analysis assumes constant neuronal activation across this period. However, fully characterizing the BOLD signal during the Cue and Fixation phases and refining the model to account for any temporal inhomogeneity in the signal falls outside the scope of this study. We recommend using one dataset to characterize the signal and a separate, independent dataset to run the model. This enhanced model should build off of the Saturated model with an aim to continue to

capture all relevant components of the trial, ensuring sufficient flexibility to accommodate variability likely present in the signal.

1 Conclusion

In task fMRI time series models, collinearity between tasks regressors may reduce statistical power for contrasts of interest. Researchers use different strategies to reduce collinearity such as omitting regressors and using impulse regressors. Using simulations and real-world analyses of the ABCD MID fMRI task we illustrated how the omission of task-relevant components, using impulse regressors and omitting response times impacts statistical validity by introducing bias into adjacent task component estimates. For the MID task, and other fMRI tasks where trials consist of many rapidly presented stimuli without interspersed fixation, we propose a Saturated time series model that models each task component. This model increases statistical validity and, in the case of the ABCD MID task, does not significantly increasing collinearity. When analyzing data, we recommend that researchers avoid omitting task-relevant regressors that are not true "rest" periods, and estimate variance inflation for regressors and contrast estimates for subject-level models. Additionally we provide and improved general VIF calculation, the contrast VIF (cVIF), that can be used to assess the impact of condition collinearity on *contrasts* of parameter estimates.

A How VIFS can be misleading

The traditional VIF estimate applies only to single regression parameters in a model and is defined as $\frac{1}{1-R^2}$, where R^2 is the coefficient of determination from regressing the regressor of interest on all other regressors in the model. As detailed in the main text, this corresponds to the ratio of a parameter's variance in the best-case scenario (no correlation with other regressors) to the worst-case scenario (actual correlations between regressors).

There are two shortcomings to this estimate of VIF. First, it cannot be used to estimate the VIF for contrasts of parameters. Second, the correlation between regressors may not reflect the true underlying correlation of the conditions as the example shown here illustrates.

Our improved contrast-based VIF calculation, cVIF, can be applied to any contrast of conditions. Also, it ensures that the ratio of the two variances truly reflect a "best" case scenario with uncorrelated *conditions* and the "worst" case scenario with the true correlation between conditions. This is achieved by focusing only on models where the regressor correlations reflect the condition correlations and estimating the two variance separately, instead of relying on the short cut of using a regression R^2 as the traditional VIF does.

The below example, also reflected in Panel A of Figure 1, illustrates how a correlation between two stimuli, A and B can be masked from a traditional VIF calculation since the correlation between A and B is not captured by the correlation between the model regressors due to the use of orthogonal regressors. We then illustrate how our cVIF calculation captures the impact of correlation between conditions on the variance of the contrast.

We use a simple fMRI task with two conditions, A and B, and the goal is to test the condition difference. The models mirror the example in the first panel of Figure 1. Define Model 1 as,

$$Y = \beta_A X_A + \beta_B X_B + \epsilon,$$

where X_A is a regressor that models condition A versus baseline and X_B models condition B versus baseline, and define Model 2 as,

$$Y = \beta_{A-B} X_{A-B} + \beta_{trials} X_{trials} + \epsilon,$$

where X_{A-B} is a parametrically modulated regressor that models the difference between the A and B conditions coded using 1's for A and -1's for B (modulation values are then mean centered) and the X_{trials} regressor represents the average activation across all trials. Due to the centering of the modulation values, X_{trials} is orthogonal to X_{A-B} since the dot product of the modulation values of X_{A-B} (centered 1/-1 values) with the modulation values of X_{trials} (all 1s, since the regressor is not modulated) is 0. The test statistics for

the A/B comparison will be identical from these two models, as such, the impact of condition correlations on $Var(\hat{\beta}_A - \hat{\beta}_B)$ from Model 1 will be the same as the impact on $Var(\hat{\beta}_{A-B})$ from Model 2, which is illustrated in the left panel of Figure A for a range of condition correlations. Assume the regressors and outcome have been centered and scaled by the standard deviation/N-1 where N is the number of observations. This obviates the need for an intercept in the model and ensures that the off-diagonal entry of the matrix $X'X$ is the correlation between the two regressors in the model, $R_{X_{AB}} = Cor(X_A, X_B)$.

Focusing on Model 1, it follows that

$$\frac{1}{\sigma^2} Var \begin{pmatrix} \hat{\beta}_A \\ \hat{\beta}_B \end{pmatrix} = (X'X)^{-1} = \begin{pmatrix} 1 & -R_{X_{AB}} \\ -R_{X_{AB}} & 1 \end{pmatrix} \frac{1}{1 - R^2_{X_{AB}}},$$

and so the worst case scenario variance, accounting for $R_{X_{AB}}$, the correlation between the A and B conditions is,

$$\begin{aligned} Var_{worst}(\hat{\beta}_A - \hat{\beta}_B) &= Var(\hat{\beta}_A) + Var(\hat{\beta}_B) - 2Cov(\hat{\beta}_A, \hat{\beta}_B) \\ &= \frac{1}{1 - R^2_{X_{AB}}}(1 + 1 + 2R_{X_{AB}}) \\ &= \frac{2}{1 - R_{X_{AB}}}. \end{aligned} \tag{1}$$

Our proposed cVIF calculation explicitly estimates the ratio of the contrast variance in the "worst" versus "best" case scenario. The "best-case" scenario for $Var(\hat{\beta}_A - \hat{\beta}_B)$ would then be if the conditions were not correlated which is when $R_{X_{AB}} = 0$. This yields $Var_{best}(\hat{\beta}_A - \hat{\beta}_B) = 2$, giving a contrast VIF of

$$cVIF_{\hat{\beta}_A - \hat{\beta}_B} = \frac{Var_{worst}(\hat{\beta}_A - \hat{\beta}_B)}{Var_{best}(\hat{\beta}_A - \hat{\beta}_B)} = \frac{\frac{2}{1 - R_{X_{AB}}}}{2} \frac{1}{1 - R_{X_{AB}}}. \tag{2}$$

For Model 2, the traditional VIF for $\hat{\beta}_{A-B}$, based on the the R^2 from the regression modeling X_{A-B} as a function of X_{trials}, takes on a value of 1 regardless of the correlation between the conditions since the regressors are uncorrelated and and do not reflect the correlations between conditions. Figure A.6 shows the relationship between $Var(\hat{\beta}_A - \hat{\beta}_B)$ and the correlation between conditions, which is reflected in $R_{X_{AB}}$, illustrating that only positive correlations between conditions has a negative impact on the condition difference. The right hand panel shows how the the cVIF estimate proposed here, $cVIF_{\hat{\beta}_A - \hat{\beta}_B}$ mirrors the variance trend correctly indicating when the variance is inflated. The cVIFs from Model 1 corresponding to the individual condition effect estimates and are only helpful in identifying collinearity when conditions are positively correlated and can misidentify high collinearity when the conditions are negatively correlated.

The traditional VIFs from Model 2 are uninformative. We recommend using the general cVIF calculation for contrasts defined in Appendix B along with the properly parameterized model.

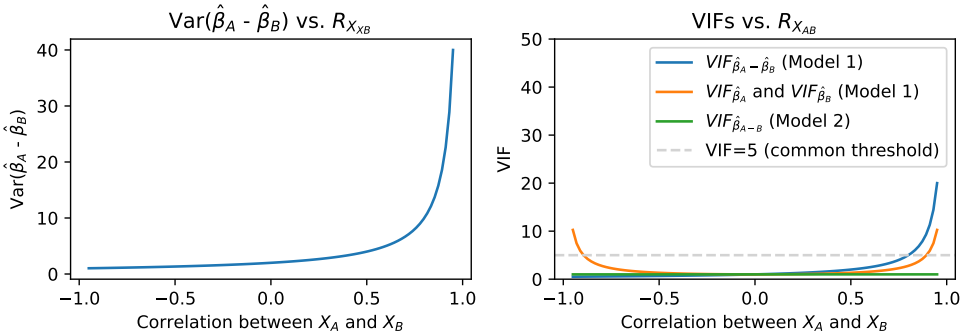

Figure A.6: VIF illustration. The left panel shows how increasing positive correlations between conditions inflates the variance of the estimated contrast comparing conditions. The right panel compares our proposed contrast-based cVIF using Model 1 (blue, see Appendix B for the general algorithm) to using the traditional, condition-specific cVIFs from Model 1 (orange) and the traditional VIF from Model 2 (green), which is uninformative since it is constant due to its inability to reflect the correlation between the A/B conditions.

B cVIF: VIF calculation for contrasts of parameter estimates

A requirement for using this cVIF calculation for contrasts of parameter estimates is that each condition must be modeled within a separate regressor, as in Model 1 of the illustration in Appendix A. Parametrically modulated regressors should only be used if there are more than 2 levels for the modulator and in the case that there are only 2 levels, they can be equivalently expressed using separate regressors for each level. The design matrix entered into the calculation should omit the intercept, but mean center each regressor. This is an important step to ensure the correlations (not the dot products) between all pairs of regressors is set to 0 in the "best" case scenario. Additionally, it is useful to divide each regressor by its standard deviation as this may stabilize the estimate of the matrix inverse. In summary, X consists of centered and scaled regressors omitting the intercept term.

To motivate the calculation, for the linear regression specified by

$$Y = X\beta + \epsilon,$$

where Y is a vector (e.g., the BOLD time series), X is the design matrix (the centered regressors described above), β is a vector of parameters and ϵ is the error term, the variance of a contrast, row vector c, of

parameter estimates is given by

$$\mathrm{Var}\left(c\hat{\beta}\right) = c\left(X'X\right)^{-1} c'\sigma^2,$$

where σ^2 is the residual variance. The cVIF estimate is a ratio of this variance under two conditions: the true condition correlations (the equation as given) versus no correlation between conditions (set all off-diagonal elements of $X'X$ to zero). Note the residual variance term drops out in the ratio, so this calculation is not dependent upon the data, Y. Given this, general cVIF calculation is defined by

$$cVIF_{c\hat{\beta}} = \frac{c(X'X)^{-1}c'}{c(X'X \circ I_N)^{-1}c'},$$

where the $\circ$ operator represents the element-wise matrix multiplier and I_N is an $N \times N$ identity matrix, where N is the number of observations. The multiplication with I_N in the denominator sets the correlation between all conditions to 0.

A Python implementation of this algorithm can be found at `https://github.com/jmumford/vif_contrasts`.

References

A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8, 2014. ISSN 1662-5196. URL https://www.frontiersin.org/articles/10.3389/fninf.2014.00014.

M. M. Andrews, S. A. Meda, A. D. Thomas, M. N. Potenza, J. H. Krystal, P. Worhunsky, M. C. Stevens, S. O'Malley, G. A. Book, B. Reynolds, and G. D. Pearlson. Individuals family history positive for alcoholism show functional magnetic resonance imaging differences in reward sensitivity that are related to impulsivity factors. *Biological Psychiatry*, 69(7):675–683, 2011. ISSN 0006-3223. doi: 10.1016/j.biopsych.2010.09.049. URL https://www.sciencedirect.com/science/article/pii/S0006322310010188.

J. Antonakis, S. Bendahan, P. Jacquart, and R. Lalive. On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21(6):1086–1120, Dec. 2010. ISSN 10489843. doi: 10.1016/j.leaqua.2010.10.010.

I. M. Balodis and M. N. Potenza. Anticipatory Reward Processing in Addicted Populations: A Focus on the Monetary Incentive Delay Task. *Biological Psychiatry*, 77(5):434–444, Mar. 2015. ISSN 0006-3223. doi: 10.1016/j.biopsych.2014.08.020. URL http://www.sciencedirect.com/science/article/pii/S0006322314006520.

I. M. Balodis, H. Kober, P. D. Worhunsky, M. C. Stevens, G. D. Pearlson, and M. N. Potenza. Diminished frontostriatal activity during processing of monetary rewards and losses in pathological gambling. *Biological Psychiatry*, 71(8):749–757, 2012. ISSN 0006-3223. doi: 10.1016/j.biopsych.2012.01.006. URL https://www.sciencedirect.com/science/article/pii/S0006322312000443.

O. Bartra, J. T. McGuire, and J. W. Kable. The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, 76:412–427, 2013. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2013.02.063.

A. Beck, F. Schlagenhauf, T. Wüstenberg, J. Hein, T. Kienast, T. Kahnt, K. Schmack, C. Hägele, B. Knutson, A. Heinz, and J. Wrase. Ventral striatal activation during reward anticipation correlates with impulsivity in alcoholics. *Biological Psychiatry*, 66(8):734–742, 2009. ISSN 0006-3223. doi: 10.1016/j.biopsych.2009.04.035. URL https://www.sciencedirect.com/science/article/pii/S0006322309005678.

J. M. Bjork. The ups and downs of relating nondrug reward activation to substance use risk in adolescents. *Current Addiction Reports*, 7(3):421–429, 2020. ISSN 2196-2952. doi: 10.1007/s40429-020-00327-7. URL https://doi.org/10.1007/s40429-020-00327-7.

J. M. Bjork, A. R. Smith, G. Chen, and D. W. Hommer. Psychosocial problems and recruitment of incentive neurocircuitry: Exploring individual differences in healthy adolescents. *Developmental Cognitive Neuroscience*, 1(4):570–577, 2011. ISSN 1878-9293. doi: 10.1016/j.dcn.2011.07.005. URL `http://www.sciencedirect.com/science/article/pii/S1878929311000703`.

C. Büchel, J. Peters, T. Banaschewski, A. L. W. Bokde, U. Bromberg, P. J. Conrod, H. Flor, D. Papadopoulos, H. Garavan, P. Gowland, A. Heinz, H. Walter, B. Ittermann, K. Mann, J.-L. Martinot, M.-L. Paillère-Martinot, F. Nees, T. Paus, Z. Pausova, L. Poustka, M. Rietschel, T. W. Robbins, M. N. Smolka, J. Gallinat, G. Schumann, and B. Knutson. Blunted ventral striatal responses to anticipated rewards foreshadow problematic drug use in novelty-seeking adolescents. *Nature Communications*, 8(1):1–11, Feb. 2017. ISSN 2041-1723. doi: 10.1038/ncomms14140. URL `https://www.nature.com/articles/ncomms14140`.

V. Calhoun, M. Stevens, G. Pearlson, and K. Kiehl. fMRI analysis with the general linear model: Removal of latency-induced amplitude bias by incorporation of hemodynamic derivative terms. *NeuroImage*, 22(1): 252–257, May 2004. ISSN 10538119. doi: 10.1016/j.neuroimage.2003.12.029.

B. Casey, T. Cannonier, M. I. Conley, A. O. Cohen, D. M. Barch, M. M. Heitzeg, M. E. Soules, T. Teslovich, D. V. Dellarco, H. Garavan, C. A. Orr, T. D. Wager, M. T. Banich, N. K. Speer, M. T. Sutherland, M. C. Riedel, A. S. Dick, J. M. Bjork, K. M. Thomas, B. Chaarani, M. H. Mejia, D. J. Hagler, M. Daniela Cornejo, C. S. Sicat, M. P. Harms, N. U. F. Dosenbach, M. Rosenberg, E. Earl, H. Bartsch, R. Watts, J. R. Polimeni, J. M. Kuperman, D. A. Fair, A. M. Dale, and ABCD Imaging Acquisition Workgroup. The adolescent brain cognitive development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, 32, 2018. ISSN 1878-9307. doi: 10.1016/j.dcn.2018.03.001. URL `https://pubmed.ncbi.nlm.nih.gov/29567376/`.

B. Chaarani, S. Hahn, N. Allgaier, S. Adise, M. M. Owens, A. C. Juliano, D. K. Yuan, H. Loso, A. Ivanciu, M. D. Albaugh, J. Dumas, S. Mackey, J. Laurent, M. Ivanova, D. J. Hagler, M. D. Cornejo, S. Hatton, A. Agrawal, L. Aguinaldo, L. Ahonen, W. Aklin, A. P. Anokhin, J. Arroyo, S. Avenevoli, D. Babcock, K. Bagot, F. C. Baker, M. T. Banich, D. M. Barch, H. Bartsch, A. Baskin-Sommers, J. M. Bjork, D. Blachman-Demner, M. Bloch, R. Bogdan, S. Y. Bookheimer, F. Breslin, S. Brown, F. J. Calabro, V. Calhoun, B. J. Casey, L. Chang, D. B. Clark, C. Cloak, R. T. Constable, K. Constable, R. Corley, L. B. Cottler, S. Coxe, R. K. Dagher, A. M. Dale, M. Dapretto, R. Delcarmen-Wiggins, A. S. Dick, E. K. Do, N. U. F. Dosenbach, G. J. Dowling, S. Edwards, T. M. Ernst, D. A. Fair, C. C. Fan, E. Feczko, S. W. Feldstein-Ewing, P. Florsheim, J. J. Foxe, E. G. Freedman, N. P. Friedman, S. Friedman-Hill, B. F. Fuemmeler, A. Galvan, D. G. Gee, J. Giedd, M. Glantz, P. Glaser, J. Godino, M. Gonzalez, R. Gonzalez,

S. Grant, K. M. Gray, F. Haist, M. P. Harms, S. Hawes, A. C. Heath, S. Heeringa, M. M. Heitzeg, R. Hermosillo, M. M. Herting, J. M. Hettema, J. K. Hewitt, C. Heyser, E. Hoffman, K. Howlett, R. S. Huber, M. A. Huestis, L. W. Hyde, W. G. Iacono, M. A. Infante, O. Irfanoglu, A. Isaiah, S. Iyengar, J. Jacobus, R. James, B. Jean-Francois, T. Jernigan, N. R. Karcher, A. Kaufman, B. Kelley, B. Kit, A. Ksinan, J. Kuperman, A. R. Laird, C. Larson, K. LeBlanc, C. Lessov-Schlagger, N. Lever, D. A. Lewis, K. Lisdahl, A. R. Little, M. Lopez, M. Luciana, B. Luna, P. A. Madden, H. H. Maes, C. Makowski, A. T. Marshall, M. J. Mason, J. Matochik, B. D. McCandliss, E. McGlade, I. Montoya, G. Morgan, A. Morris, C. Mulford, P. Murray, B. J. Nagel, M. C. Neale, G. Neigh, A. Nencka, A. Noronha, S. J. Nixon, C. E. Palmer, V. Pariyadath, M. P. Paulus, W. E. Pelham, D. Pfefferbaum, C. Pierpaoli, A. Prescot, D. Prouty, L. I. Puttler, N. Rajapaske, K. M. Rapuano, G. Reeves, P. F. Renshaw, M. C. Riedel, P. Rojas, M. de la Rosa, M. D. Rosenberg, M. J. Ross, M. Sanchez, C. Schirda, D. Schloesser, J. Schulenberg, K. J. Sher, C. Sheth, P. D. Shilling, W. K. Simmons, E. R. Sowell, N. Speer, M. Spittel, L. M. Squeglia, C. Sripada, J. Steinberg, C. Striley, M. T. Sutherland, J. Tanabe, S. F. Tapert, W. Thompson, R. L. Tomko, K. A. Uban, S. Vrieze, N. E. Wade, R. Watts, S. Weiss, B. A. Wiens, O. D. Williams, A. Wilbur, D. Wing, D. Wolff-Hughes, R. Yang, D. A. Yurgelun-Todd, R. A. Zucker, A. Potter, and H. P. Garavan. Baseline brain function in the preadolescents of the ABCD study. *Nature Neuroscience*, pages 1–11, 2021. ISSN 1546-1726. doi: 10.1038/s41593-021-00867-9. URL `https://www.nature.com/articles/s41593-021-00867-9`.

M. I. Demidenko, J. A. Mumford, and R. A. Poldrack. Impact of analytic decisions on test–retest reliability of individual and group estimates in functional magnetic resonance imaging: A multiverse analysis using the monetary incentive delay task. *Imaging Neuroscience*, 2:1–26, Sept. 2024. ISSN 2837-6056. doi: 10.1162/imag_a_00262.

M. L. Elliott, A. R. Knodt, D. Ireland, M. L. Morris, R. Poulton, S. Ramrakha, M. L. Sison, T. E. Moffitt, A. Caspi, and A. R. Hariri. What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, 31(7):792–806, July 2020. ISSN 0956-7976, 1467-9280. doi: 10.1177/0956797620916786.

O. Esteban, C. J. Markiewicz, M. Goncalves, C. Provins, J. D. Kent, E. DuPre, T. Salo, R. Ciric, B. Pinsard, R. W. Blair, R. A. Poldrack, and K. J. Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI, Sept. 2022. URL `https://zenodo.org/record/7117719`.

E. Feczko, G. Conan, S. Marek, B. Tervo-Clemmens, M. Cordova, O. Doyle, E. Earl, A. Perrone, D. Sturgeon, R. Klein, G. Harman, D. Kilamovich, R. Hermosillo, O. Miranda-Dominguez, A. Adebimpe, M. Bertolero, M. Cieslak, S. Covitz, T. Hendrickson, A. C. Juliano, K. Snider, L. A. Moore, J. Uriartel, A. M. Graham,

F. Calabro, M. D. Rosenberg, K. M. Rapuano, B. J. Casey, R. Watts, D. Hagler, W. K. Thompson, T. E. Nichols, E. Hoffman, B. Luna, H. Garavan, T. D. Satterthwaite, S. F. Ewing, B. Nagel, N. U. F. Dosenbach, and D. A. Fair. Adolescent Brain Cognitive Development (ABCD) Community MRI Collection and Utilities. *bioRxiv*, 2021. doi: 10.1101/2021.07.09.451638. URL `https://www.biorxiv.org/content/10.1101/2021.07.09.451638v1`.

K. A. Garrison, S. W. Yip, I. M. Balodis, K. M. Carroll, M. N. Potenza, and S. Krishnan-Sarin. Reward-related frontostriatal activity and smoking behavior among adolescents in treatment for smoking cessation. *Drug and Alcohol Dependence*, 177:268–276, 2017. ISSN 1879-0046. doi: 10.1016/j.drugalcdep.2017.03.035.

W. H. Greene. *Econometric Analysis*. Pearson Education Inc., New York, 5th edition, 2003. ISBN 0-13-066189-9.

D. N. Greve and B. Fischl. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48:63–72, 2009. doi: 10.1016/j.neuroimage.2009.06.060.

D. J. Hagler, S. Hatton, M. D. Cornejo, C. Makowski, D. A. Fair, A. S. Dick, M. T. Sutherland, B. J. Casey, D. M. Barch, M. P. Harms, R. Watts, J. M. Bjork, H. P. Garavan, L. Hilmer, C. J. Pung, C. S. Sicat, J. Kuperman, H. Bartsch, F. Xue, M. M. Heitzeg, A. R. Laird, T. T. Trinh, R. Gonzalez, S. F. Tapert, M. C. Riedel, L. M. Squeglia, L. W. Hyde, M. D. Rosenberg, E. A. Earl, K. D. Howlett, F. C. Baker, M. Soules, J. Diaz, O. R. de Leon, W. K. Thompson, M. C. Neale, M. Herting, E. R. Sowell, R. P. Alvarez, S. W. Hawes, M. Sanchez, J. Bodurka, F. J. Breslin, A. S. Morris, M. P. Paulus, W. K. Simmons, J. R. Polimeni, A. van der Kouwe, A. S. Nencka, K. M. Gray, C. Pierpaoli, J. A. Matochik, A. Noronha, W. M. Aklin, K. Conway, M. Glantz, E. Hoffman, R. Little, M. Lopez, V. Pariyadath, S. R. Weiss, D. L. Wolff-Hughes, R. DelCarmen-Wiggins, S. W. Feldstein Ewing, O. Miranda-Dominguez, B. J. Nagel, A. J. Perrone, D. T. Sturgeon, A. Goldstone, A. Pfefferbaum, K. M. Pohl, D. Prouty, K. Uban, S. Y. Bookheimer, M. Dapretto, A. Galvan, K. Bagot, J. Giedd, M. A. Infante, J. Jacobus, K. Patrick, P. D. Shilling, R. Desikan, Y. Li, L. Sugrue, M. T. Banich, N. Friedman, J. K. Hewitt, C. Hopfer, J. Sakai, J. Tanabe, L. B. Cottler, S. J. Nixon, L. Chang, C. Cloak, T. Ernst, G. Reeves, D. N. Kennedy, S. Heeringa, S. Peltier, J. Schulenberg, C. Sripada, R. A. Zucker, W. G. Iacono, M. Luciana, F. J. Calabro, D. B. Clark, D. A. Lewis, B. Luna, C. Schirda, T. Brima, J. J. Foxe, E. G. Freedman, D. W. Mruzek, M. J. Mason, R. Huber, E. McGlade, A. Prescot, P. F. Renshaw, D. A. Yurgelun-Todd, N. A. Allgaier, J. A. Dumas, M. Ivanova, A. Potter, P. Florsheim, C. Larson, K. Lisdahl, M. E. Charness, B. Fuemmeler, J. M. Hettema, H. H. Maes, J. Steinberg, A. P. Anokhin, P. Glaser, A. C. Heath, P. A. Madden, A. Baskin-Sommers, R. T. Constable, S. J. Grant, G. J. Dowling, S. A. Brown, T. L. Jernigan, and A. M. Dale. Image processing and

analysis methods for the adolescent brain cognitive development study. *NeuroImage*, 202:116091, 2019. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2019.116091.

M. Jenkinson, P. Bannister, M. Brady, and S. Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841, Oct. 2002. doi: 10.1006/nimg.2002.1132.

J. E. Joseph, X. Zhu, D. Lynam, and T. H. Kelly. Modulation of meso-limbic reward processing by motivational tendencies in young adolescents and adults. *NeuroImage*, 129:40–54, Apr. 2016. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.12.005. URL `https://www.sciencedirect.com/science/article/pii/S1053811915011180`.

B. Knutson and E. G. Wimmer. Splitting the difference. *Annals of the New York Academy of Sciences*, 1104(1):54–69, 2007. ISSN 1749-6632. doi: 10.1196/annals.1390.020. URL `https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1196/annals.1390.020`.

B. Knutson, C. M. Adams, G. W. Fong, and D. Hommer. Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience*, 21(16):RC159–RC159, 2001. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.21-16-j0002.2001. URL `http://www.jneurosci.org/content/21/16/RC159`.

B. Knutson, G. W. Fong, S. M. Bennett, C. M. Adams, and D. Hommer. A region of mesial prefrontal cortex tracks monetarily rewarding outcomes: characterization with rapid event-related fMRI. *NeuroImage*, 18(2):263–272, 2003. ISSN 1053-8119. doi: 10.1016/S1053-8119(02)00057-5. URL `https://www.sciencedirect.com/science/article/pii/S1053811902000575`.

B. Knutson, J. Taylor, M. Kaufman, R. Peterson, and G. Glover. Distributed Neural Representation of Expected Value. *Journal of Neuroscience*, 25(19):4806–4812, May 2005. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0642-05.2005. URL `https://www.jneurosci.org/content/25/19/4806`.

B. Knutson, K. Katovich, and G. Suri. Inferring affect from fMRI data. *Trends in Cognitive Sciences*, 18 (8):422–428, 2014. ISSN 1879-307X. doi: 10.1016/j.tics.2014.04.006. URL `https://pubmed.ncbi.nlm.nih.gov/24835467/`.

M. A. Lindquist and T. D. Wager. Validity and power in hemodynamic response modeling: A comparison study and a new approach. *Human Brain Mapping*, 28(8):764–784, Nov. 2006. ISSN 1065-9471. doi: 10.1002/hbm.20310.

T. T. Liu, L. R. Frank, E. C. Wong, and R. B. Buxton. Detection Power, Estimation Efficiency, and Predictability in Event-Related fMRI. *NeuroImage*, 13(4):759–773, Apr. 2001. ISSN 10538119. doi: 10.1006/nimg.2000.0728.

M. E. Martz, E. M. Trucco, L. M. Cope, J. E. Hardee, J. M. Jester, R. A. Zucker, and M. M. Heitzeg. Association of marijuana use with blunted nucleus accumbens response to reward anticipation. *JAMA Psychiatry*, 73(8):838–844, 2016. ISSN 2168-622X. doi: 10.1001/jamapsychiatry.2016.1161. URL `https://jamanetwork.com/journals/jamapsychiatry/fullarticle/2532234`.

J. A. Mumford and T. E. Nichols. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *NeuroImage*, 39(1):261–268, Jan. 2008. ISSN 10538119. doi: 10.1016/j.neuroimage.2007.07.061.

J. A. Mumford, P. G. Bissett, H. M. Jones, S. Shim, J. A. H. Rios, and R. A. Poldrack. The response time paradox in functional magnetic resonance imaging analyses. *Nature Human Behaviour*, 8(2):349–360, Feb. 2024. ISSN 2397-3374. doi: 10.1038/s41562-023-01760-0. URL `https://www.nature.com/articles/s41562-023-01760-0`. Publisher: Nature Publishing Group.

D. E. Nee. fMRI replicability depends upon sufficient individual-level data. *Communications Biology*, 2(1): 130, Apr. 2019. ISSN 2399-3642. doi: 10.1038/s42003-019-0378-6.

S. Oldham, C. Murawski, A. Fornito, G. Youssef, M. Yücel, and V. Lorenzetti. The anticipation and outcome phases of reward and loss processing: A neuroimaging meta-analysis of the monetary incentive delay task. *Human Brain Mapping*, 39(8), 2018. ISSN 1097-0193. doi: 10.1002/hbm.24184.

K. T. Patel, M. C. Stevens, S. A. Meda, C. Muska, A. D. Thomas, M. N. Potenza, and G. D. Pearlson. Robust Changes in Reward Circuitry During Reward Loss in Current and Former Cocaine Users During Performance of a Monetary Incentive Delay Task. *Biological Psychiatry*, 74(7):529–537, Oct. 2013. ISSN 0006-3223. doi: 10.1016/j.biopsych.2013.04.029. URL `https://www.sciencedirect.com/science/article/pii/S0006322313004095`.

R. A. Poldrack, C. I. Baker, J. Durnez, K. J. Gorgolewski, P. M. Matthews, M. R. Munafò, T. E. Nichols, J.-B. Poline, E. Vul, and T. Yarkoni. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2):115–126, Feb. 2017. ISSN 1471-0048. doi: 10.1038/nrn.2016.167. URL `https://www.nature.com/articles/nrn.2016.167`. Number: 2 Publisher: Nature Publishing Group.

Psychology Software Tools. Timing: Timing of e-objects 22852. https://support.pstnet.com/hc/en-us/articles/115000902428-TIMING-Timing-of-E-Objects-22852, 2024. Accessed: 2024-10-25.

T. D. Satterthwaite, M. A. Elliott, R. T. Gerraty, K. Ruparel, J. Loughead, M. E. Calkins, S. B. Eickhoff, H. Hakonarson, R. C. Gur, R. E. Gur, and D. H. Wolf. An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage*, 64:240–256, Jan. 2013. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2012.08.052. URL https://www.sciencedirect.com/science/article/pii/S1053811912008609.

W. Schultz. Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80(1):1–27, 1998. ISSN 0022-3077. doi: 10.1152/jn.1998.80.1.1.

T. Srirangarajan, L. Mortazavi, T. Bortolini, J. Moll, and B. Knutson. Multi-band FMRI compromises detection of mesolimbic reward responses. *NeuroImage*, 244:118617, Dec. 2021. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2021.118617.

C. Summerfield, M. Greene, T. Wager, T. Egner, J. Hirsch, and J. Mangels. Neocortical Connectivity during Episodic Memory Formation. *PLoS Biology*, 4(5):e128, Apr. 2006. ISSN 1545-7885. doi: 10.1371/journal.pbio.0040128.

B. O. Turner, T. Santander, E. J. Paul, A. K. Barbey, and M. B. Miller. Reply to: fMRI replicability depends upon sufficient individual-level data. *Communications Biology*, 2(1):129, Apr. 2019. ISSN 2399-3642. doi: 10.1038/s42003-019-0379-5.

C. C. Wu, G. R. Samanez-Larkin, K. Katovich, and B. Knutson. Affective traits link to reliable neural markers of incentive anticipation. *NeuroImage*, 84:279–289, 2014. doi: 10.1016/j.neuroimage.2013.08.055.

W.-Y. W. Yau, J.-K. Zubieta, B. J. Weiland, P. G. Samudra, R. A. Zucker, and M. M. Heitzeg. Nucleus accumbens response to incentive stimuli anticipation in children of alcoholics: Relationships with precursive behavioral risk and lifetime alcohol use. *The Journal of Neuroscience*, 32(7):2544–2551, 2012. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.1390-11.2012. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3567451/.

Supplementary Material

S1 Accuracy of onset times in the ABCD data distribution

In May 2024, it was discovered and confirmed by the DAIRC that there was an error in the (a) Hagler et al. 2019 description of the number of calibration volumes in GE scanners and (b) the calculation of the behavioral timing files used in the task fMRI first level analyses and released with the minimally preprocessed data. First, in the initial description of the ABCD acquisition and task analysis protocol [Hagler et al., 2019], it was stated that the calibration sequence made up 16 volumes in the GE DV26, 5 in GE DV25 (one of the calibration volumes appears to comprise a collapsed 12 calibration volumes), 8 in Siemens and 8 in Philips scanners. It was inferred that the E-Prime tasks are triggered consistently across GE, Siemens and Philips scanners. However, it was discovered that the E-Prime tasks are triggered by the scanner at the start of the 5th and 16th calibration volume in the GE DV25 and GE DV26 scanners, respectively, instead of the start of the 6th and 17th volume). Meanwhile, in the Siemens/Philips scanners, the e-prime tasks are triggered correctly, starting following the end of the 8th calibration volume (i.e., start of 9th volume). As a result, the timing information in the GE data was displaced by 800 ms (1 volume). This was confirmed by analysis of the original and revised data using response-locked averaging, which showed that the offset was evident in the neural time course (Figure S1). Nonetheless, the DAIRC's reanalyses of the data following this correction resulted in only minor differences in estimated contrast beta weights for the ABCD model.

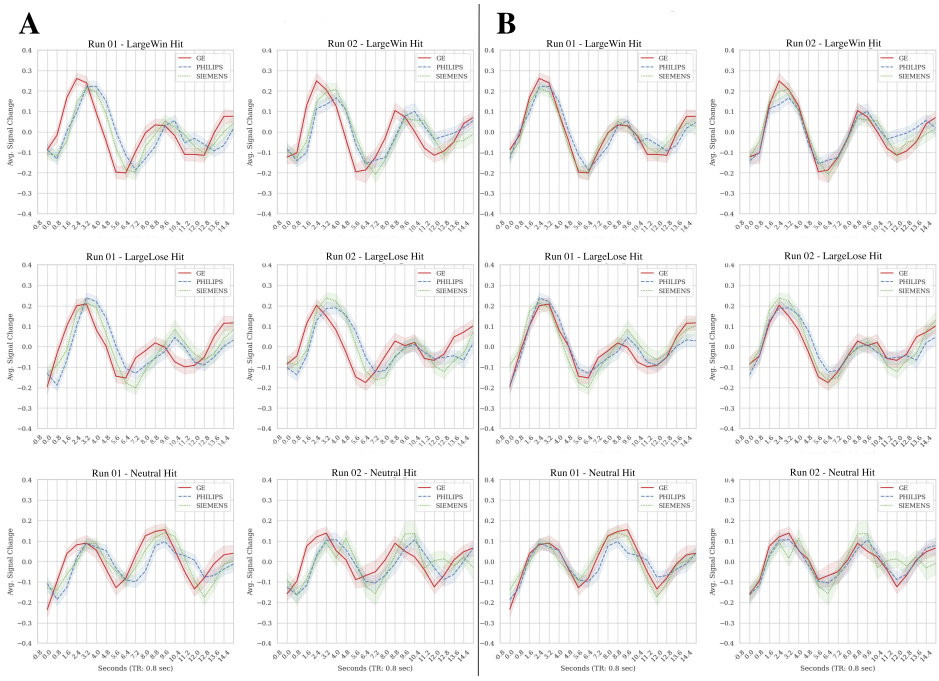

Figure S1: TR-by-TR plots locked to the Probe onset in the MID task for voxels averaged within an 8 mm radius sphere in the Left Motor Area (MNI: -38, -22, 56) for N = 50 GE, N = 50 Siemens and N = 50 Philips Baseline data for (A) Uncorrected and (B) Corrected data

S2 Design Matrices in Simulations and Data Analyses

Both the simulated and real data analyses used the design matrices shown in Figure S2, but the real data analyses included additional nuisance regressors, such as motion regressors and discrete cosine filter with 128s cut-off.

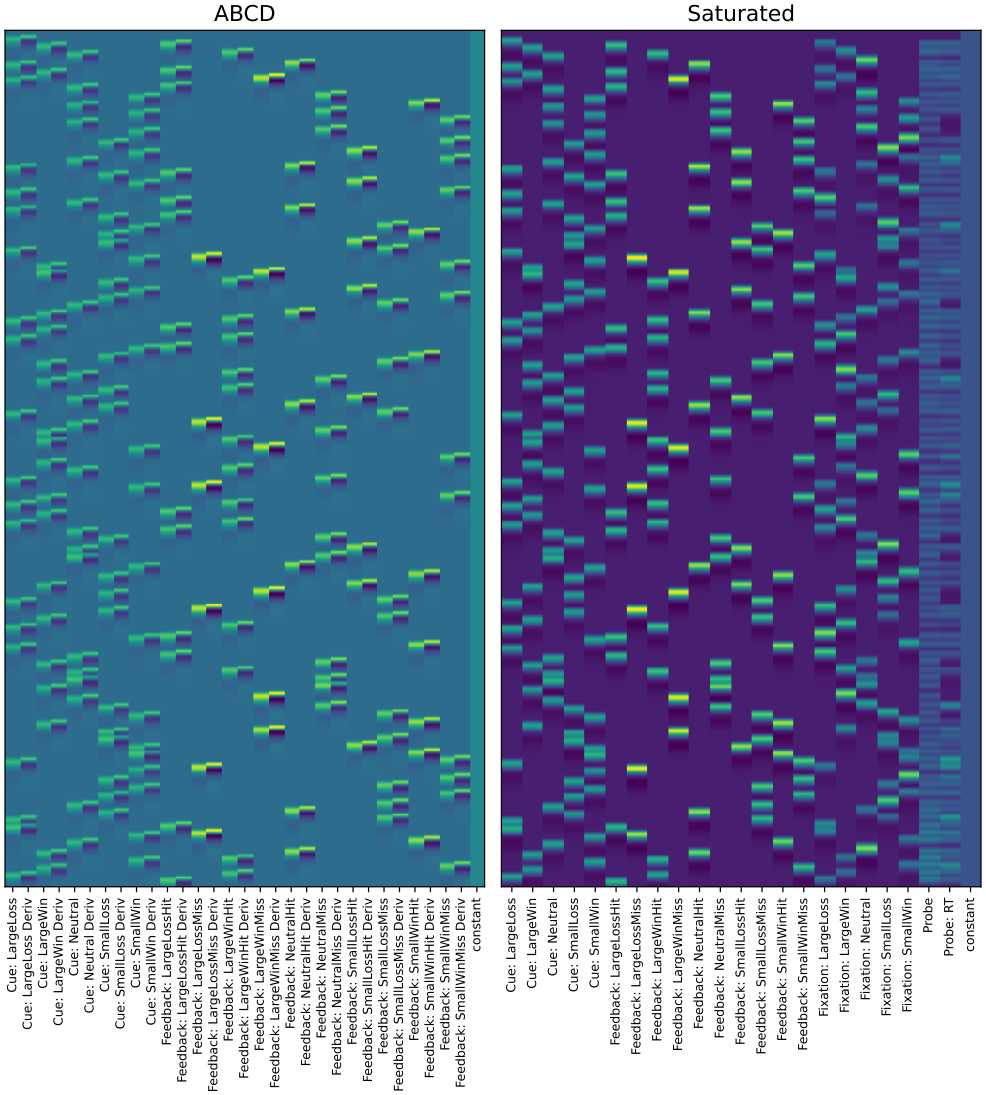

Figure S2: Design matrices used in simulated and real data analyses. The real data analysis design matrices also included cosine regressors to model low frequency drift (<0.008 Hz) and 12 motion parameters (translations, rotations and their derivatives).

S3 cVIFs and efficiencies of design matrices

The cVIFs for all regressors and contrasts used in the simulation analyses are displayed in Figure S3 and the efficiencies for the regressors that are shared between the models and the contrasts are shown in Figure S4

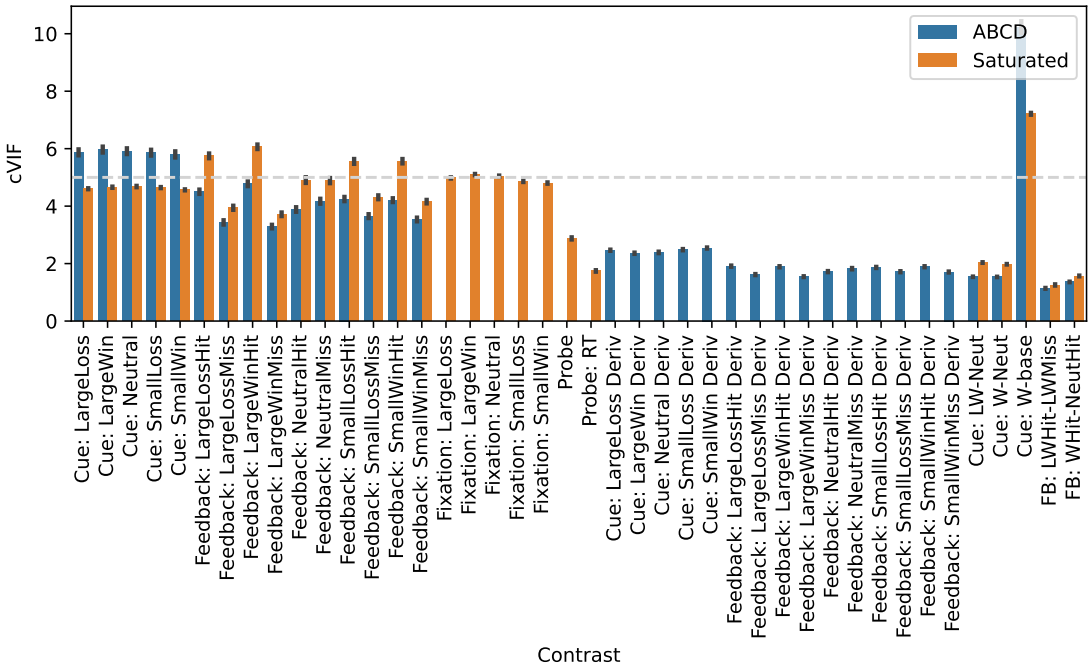

Figure S3: Variance inflation factors, using the cVIF algorithm, for all regressors and contrasts for both models.

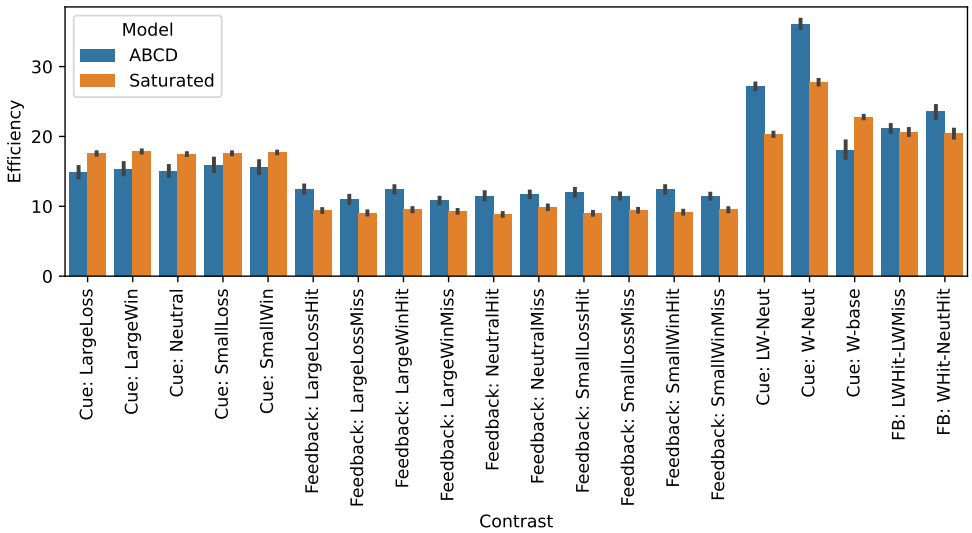

Figure S4: Efficiency estimates for all regressors that are shared by both models as well as the contrasts used in the simulation study.

S4 Stimulus durations and response times split by all three trial outcomes

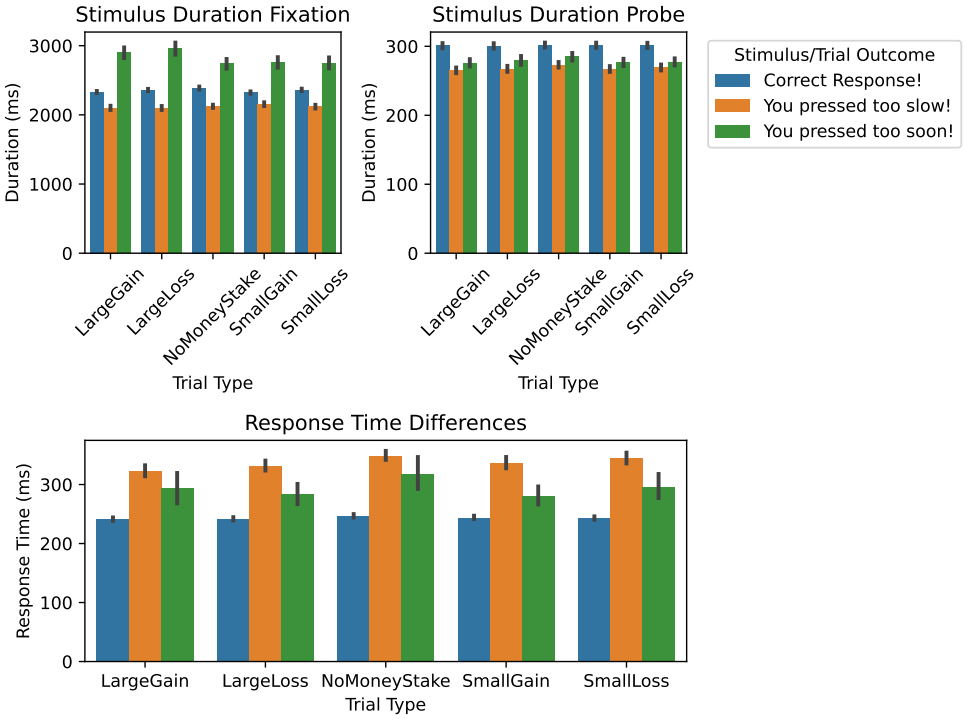

Figure S5: Stimulus durations and response times split by Correct, Too Soon and Too Slow. Note that Miss trials consist of both Too Soon and Too Slow trials, which have very different durations and RTs. On average, 57%, 36% and 7% of trials are Correct, Too Slow and Too Soon, respectively.

S5 Error rates and bias for all parameters and contrasts

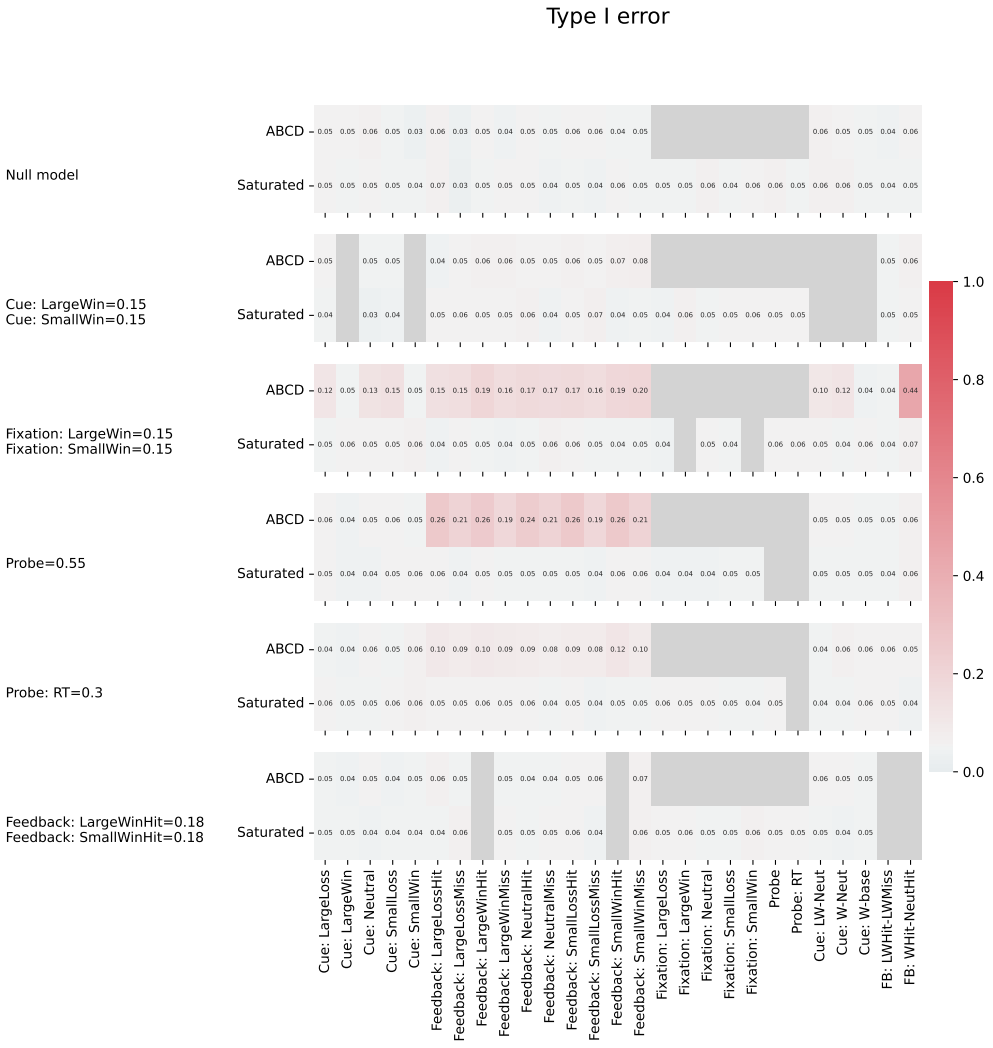

Figure S6: Type I error rates across simulation settings and models. Each row corresponds to a different model where the parameter or contrast error rate is shown for the ABCD (top) and Saturated (bottom) models, where values less than 0.05 indicate controlled Type I errors. Note the grayed out rectangles indicate that a parameter involved in that contrast was nonzero and so error rate is not estimated.

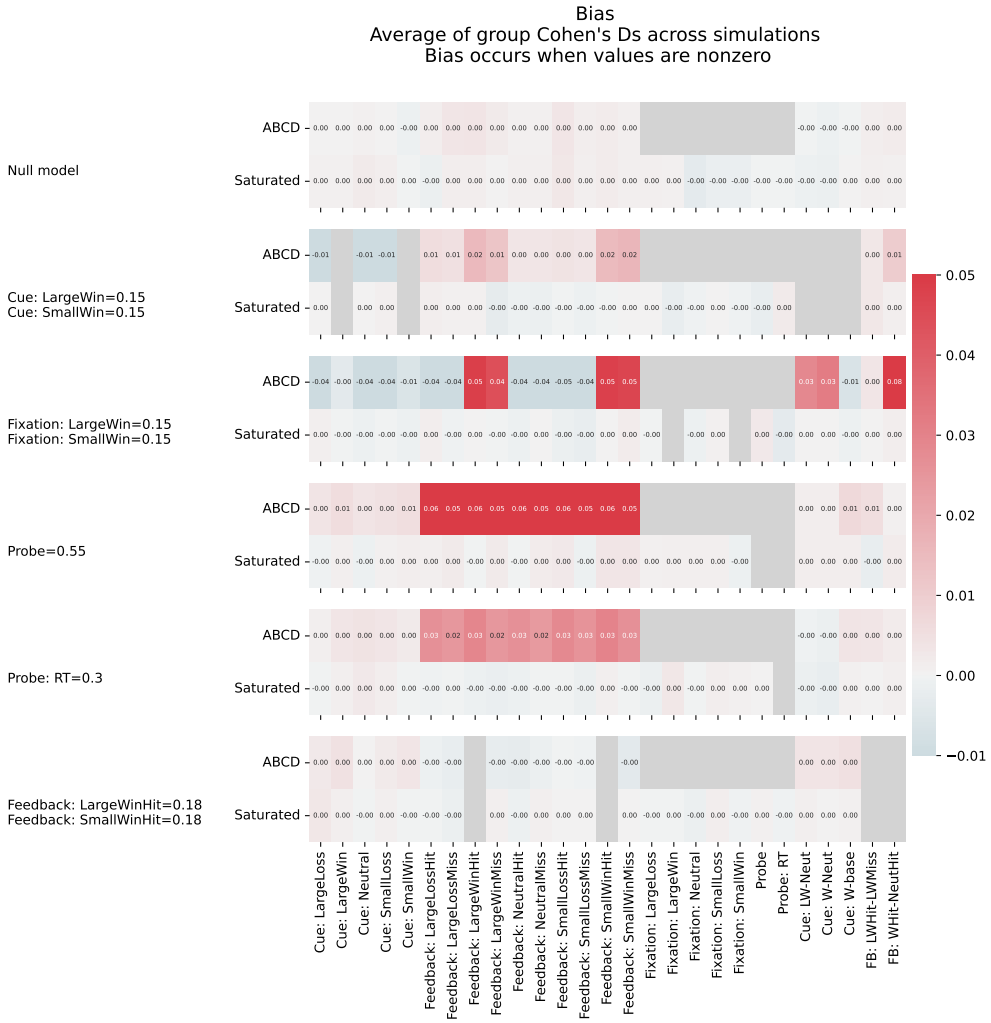

Figure S7: Bias in parameter and contrast estimates across simulation settings and models. Each row corresponds to a different model where the contrast estimate bias is shown for the ABCD (top) and Saturated (bottom) models, where 0 would indicate no bias. Note the grayed out rectangles indicate that a parameter involved in that contrast was nonzero and so bias is not estimated here.

S6 Region of Interest: Nucleus Accumbens

Binarized masks for the Nucleus Accumbens (NAc) region of interest are created using the Harvard-Oxford subcortical atlases from *Nilearn*. First, the relevant subcortical regions— left and right NAc —are extracted based on their index values in the atlas. The ROI images are then resampled to align with the space of the reference beta map (MNI152 2009cAsym). The resampled ROI images are binarized to create binary masks Figure S8.

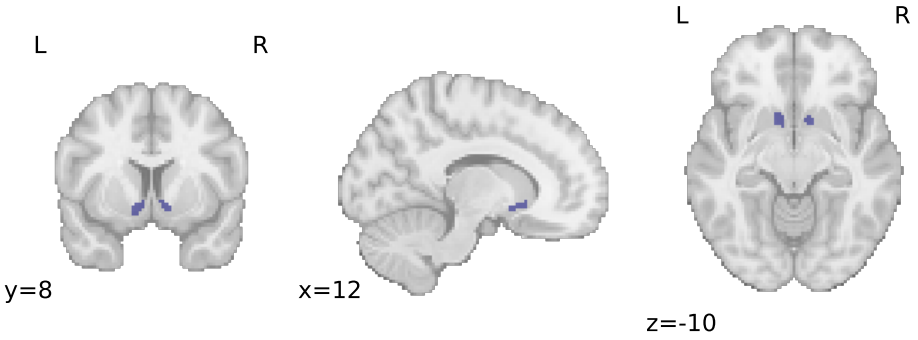

Figure S8: Harvard-Oxford subcortical Left and Right Nucleus Accumbens.

S7 Supplemental Behavioral Results

The average Probe hits (%) across Cue types are reported in Figure S11.

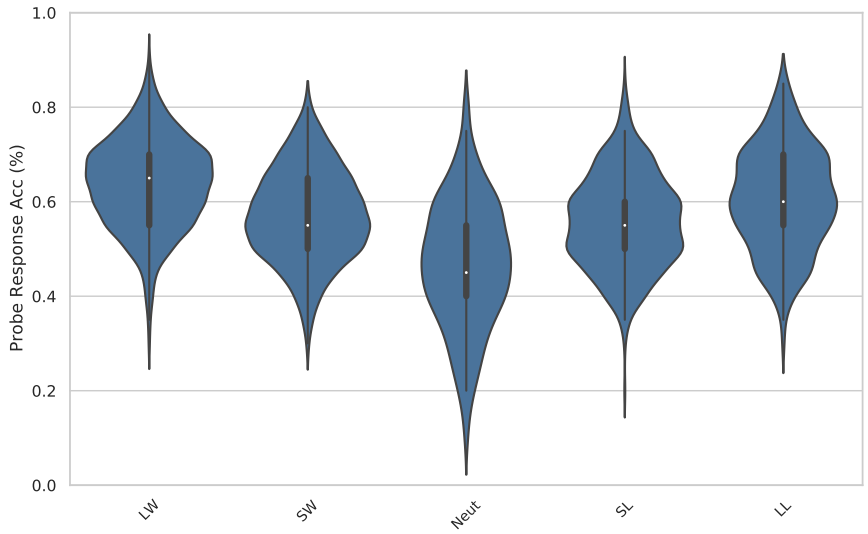

Figure S9: Probe response accuracy (%) across Cue type and B. Mean response times (RTs) across Feedback Hit/Miss conditions

S8 Supplemental Group-level Statistical Maps

In addition to the contrasts in the main text, we provide an expanded list of contrasts below in Table S1

Contrast Name	Contrast Definition
Cue: LW-Neut*	1 × Cue Large Win - 1 × Cue Neutral
Cue: W-Neut	.5 × Cue Large Win + .5 × Cue Small Win - 1 × Cue Neutral
Cue: LL-Neut*	1 × Cue Large Loss - 1 × Cue Neutral
Cue: L-Neut	.5 × Cue Large Loss + .5 × Cue Small Loss - 1 × Cue Neutral
Cue: LW-Base	1 × Cue Large Win
Fix: LW-Neut	1 × Fixation Large Win - 1 × Fixation Neutral
Fix: W-Neut	.5 × Fixation Large Win + .5 × Fixation Small Win - 1 × Fixation Neutral
Fix: LL-Neut	1 × Fixation Large Loss - 1 × Fixation Neutral
Fix: L-Neut	.5 × Fixation Large Loss + .5 × Fixation Small Loss - 1 × Fixation Neutral
Fix: LW-Base	1 × Fixation Large Win
FB: WHit-WMiss	.5 × Feedback Large Win Hit + .5 × Feedback Small Win Hit - .5 × Feedback Large Win Miss + .5 × Feedback Small Win Miss
FB: LWHit-LWMiss*	1 × Feedback Large Win Hit - 1 × Feedback Large Win Miss
FB: LHit-LMiss	.5 × Feedback Large Loss Hit + .5 × Feedback Small Loss Hit - .5 × Feedback Large Loss Miss + .5 × Feedback Small Loss Miss
FB: LLHit-LLMiss*	1 × Feedback Large Loss Hit - 1 × Feedback Large Loss Miss
FB: LWHit-NeutHit	1 × Feedback Large Win Hit - 1 × Feedback Neutral Hit
FB: LWHit-Base	1 × Feedback Large Win Hit
Probe	1 × Probe
RT	1 × RT

Table S1: Definitions of contrasts focused on for our main real data analysis results.
* Contrasts included with ABCD data releases Chaarani et al., 2021.

As mentioned in the main text, contrasts 1-4 and 11-14 are similar to those that are estimated and distributed in the ABCD consortium task fMRI derivatives Chaarani et al., 2021.

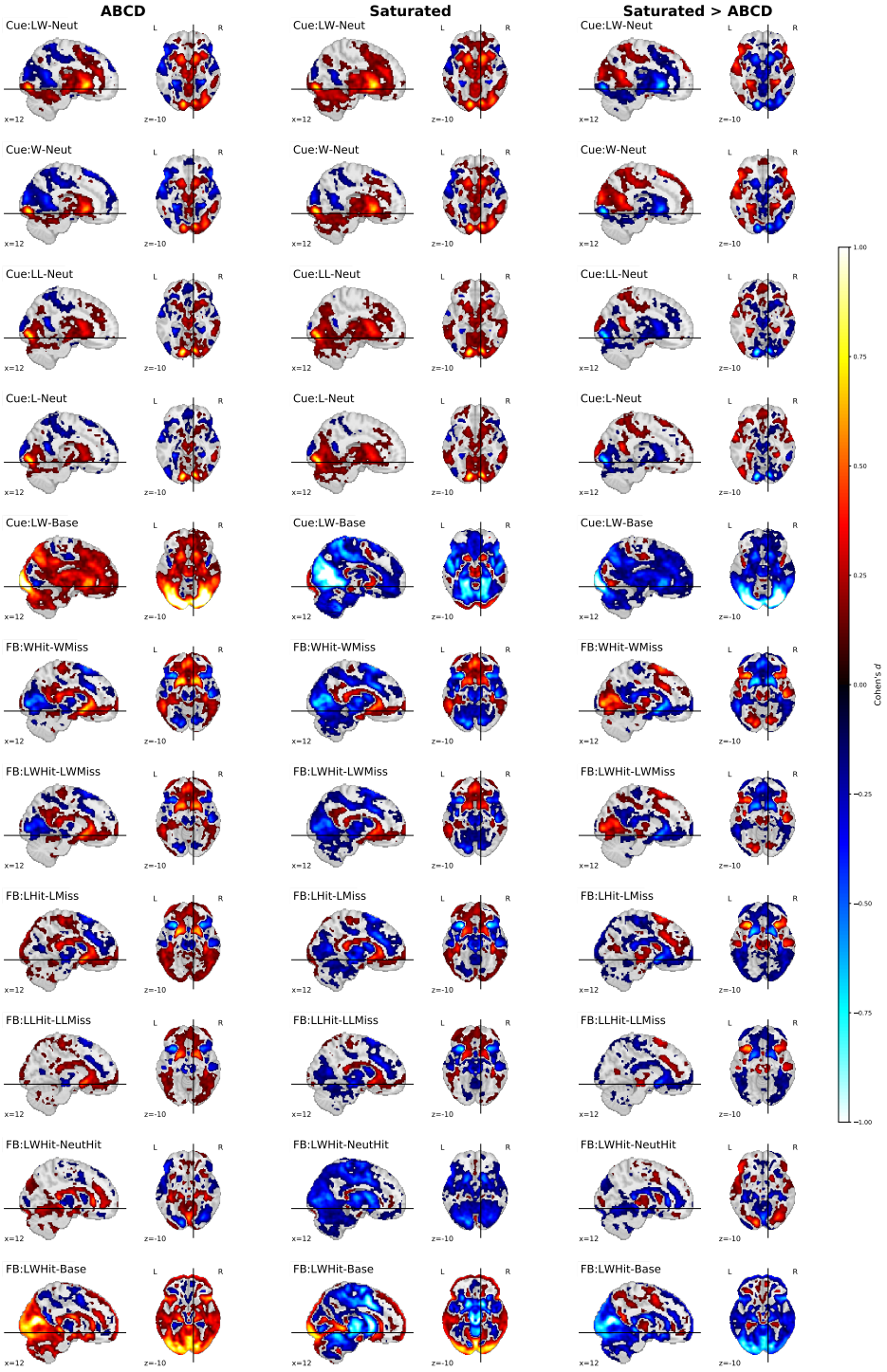

Figure S10: Cohen's d group-level activation maps (voxels selected using TFCE $p < .05$, 1000 permutations) across Cue (5) and Feedback (6) contrasts for the ABCD (shown on the left), Saturated (shown in the middle) and their paired difference (shown on the right)

The below Figure S11 illustrates the group-level effect of the Probe cue and the Probe RTs.

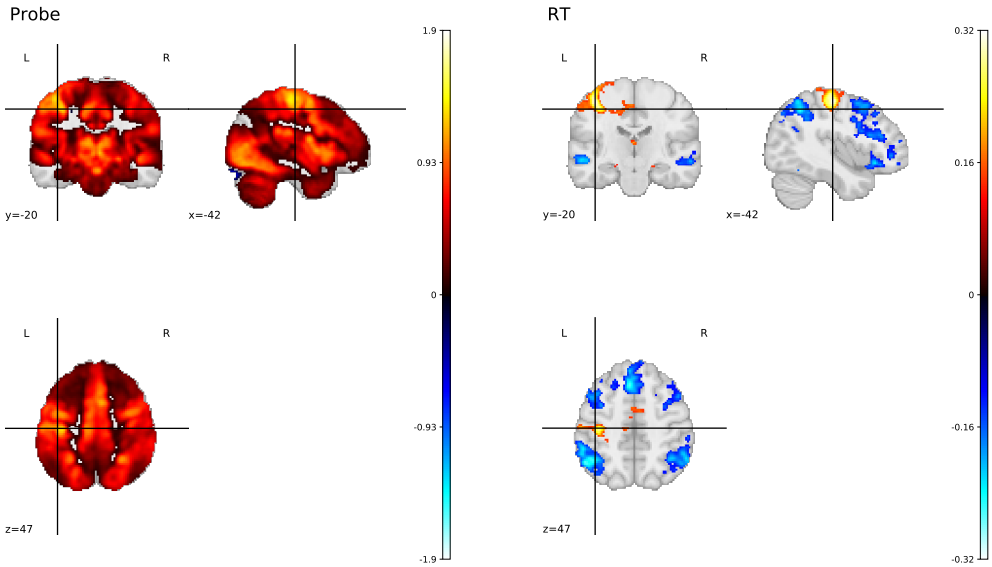

Figure S11: Group average effect (Cohen's d) of Probe and RT regressors (TFCE $p < .05$)

Figure S12 presents the group-level activity for five contrasts across the Cue and Fixation components of the MID task trials.

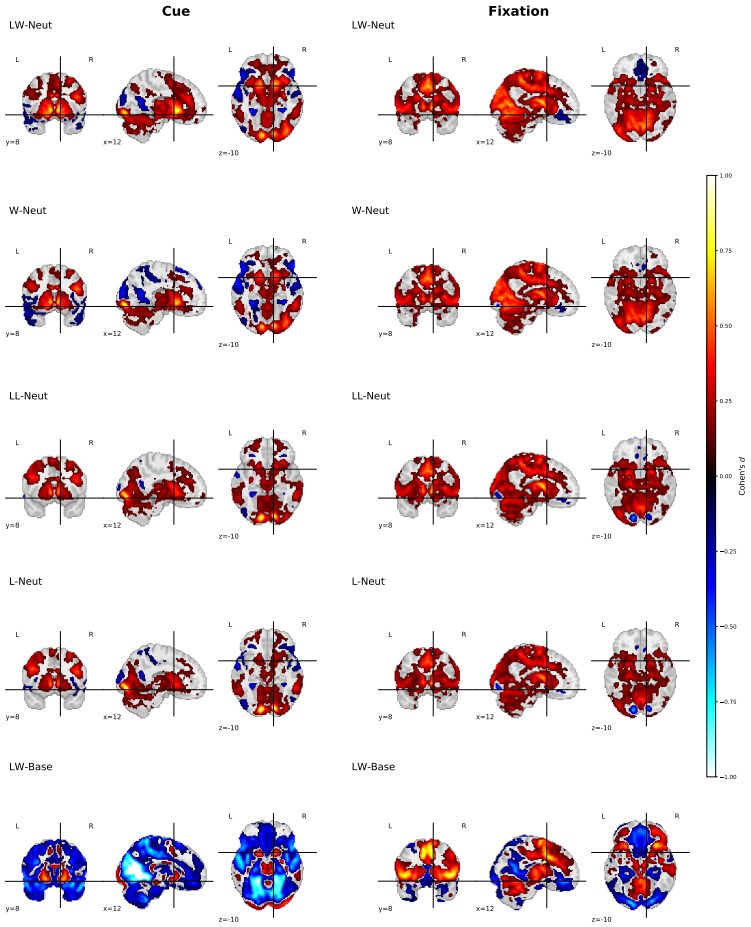

Figure S12: Cohen's d group-level activation maps (voxels selected using TFCE $p < .05$, 1000 permutations) across a Cue (5; shown on the left) and Fixation (5; shown on the right) contrasts for the Saturated model