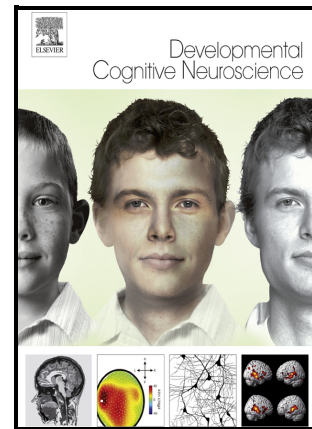


Evaluating the reliability of functional near-infrared spectroscopy data in the context of a reasoning paradigm

Patrick G. Kelly, Heather Bortfeld, Silvia A. Bunge, Keanan J. Joyner



PII: S1878-9293(26)00105-2

DOI: <https://doi.org/10.1016/j.dcn.2026.101763>

Reference: DCN101763

To appear in: *Developmental Cognitive Neuroscience*

Received date: 30 September 2025

Revised date: 12 June 2026

Accepted date: 16 June 2026

Please cite this article as: Patrick G. Kelly, Heather Bortfeld, Silvia A. Bunge and Keanan J. Joyner, Evaluating the reliability of functional near-infrared spectroscopy data in the context of a reasoning paradigm, *Developmental Cognitive Neuroscience*, (2026) doi:<https://doi.org/10.1016/j.dcn.2026.101763>

This is a PDF of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability. This version will undergo additional copyediting, typesetting and review before it is published in its final form. As such, this version is no longer the Accepted Manuscript, but it is not yet the definitive Version of Record; we are providing this early version to give early visibility of the article. Please note that Elsevier's sharing policy for the Published Journal Article applies to this version, see: <https://www.elsevier.com/about/policies-and-standards/sharing#4-published-journal-article>. Please also note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Evaluating the reliability of functional near-infrared spectroscopy data in the context of a reasoning paradigm

Patrick G. Kelly^{1,4}, Heather Bortfeld², Silvia A. Bunge^{1,3*}, & Keanan J. Joyner^{1*}

¹ Department of Psychology, University of California Berkeley, USA

² Department of Psychological Sciences, University of California Merced, USA

³ Helen Wills Neuroscience Institute, University of California Berkeley, USA

⁴ Department of Psychology, University of California Los Angeles, USA

** Joint senior authors*

Corresponding author: Patrick G. Kelly (patrickkelly@g.ucla.edu)

[Page break]

Abstract

Functional near-infrared spectroscopy (fNIRS) is a portable, motion-tolerant neuroimaging method particularly well suited for developmental and naturalistic research. To evaluate the utility of fNIRS for studying individual differences and longitudinal changes, we measured activation and functional connectivity during a relational reasoning task in young adults (N = 73). We sought to (1) establish whether fNIRS captures frontoparietal activation patterns consistent with prior fMRI studies using similar paradigms, (2) assess the effect of the amount of data (number of task blocks) on signal strength and precision, (3) assess the paradigm's measurement properties in the form of intra- and interindividual stability of activation and functional connectivity within and across testing sessions, and (4) examine whether grouping channels into anatomical regions of interest (ROIs) conferred benefits to the above. We observed robust task-evoked activation across lateral prefrontal and parietal cortices, with effect sizes on par with prior fMRI studies. Generally, we observed diminishing returns in effect size and measurement precision beyond ~7 minutes. Internal consistency and test-retest reliability varied across metrics; while they were very low for a specific task contrast, they were extremely high for functional connectivity, confirming the robustness of channel- and ROI-level connectivity as a stable marker of functional architecture. Exploratory analyses supported prior observations of lower signal quality in participants with darker skin tones and hair, underscoring the need for inclusive methodological strategies. Together, these findings highlight key design considerations for optimizing longitudinal and individual-differences research on higher-level cognition, particularly in diverse and developmentally variable populations.

Keywords (7/7):

functional near-infrared spectroscopy, psychometric, validation, test-retest reliability, development, rostrolateral prefrontal, parietal

1. Introduction

A central goal of developmental cognitive neuroscience is to understand the neural changes underpinning higher-level cognition, including reasoning. Prior work using functional magnetic resonance imaging (fMRI) has yielded insights into the neural basis of relational reasoning and its development from middle childhood onward (Dumontheil et al., 2010; Wendelken et al., 2011; 2017). However, to extend this research to younger and more diverse and representative populations and naturalistic settings, there are advantages to using alternative neuroimaging methods that are portable, versatile in terms of testing conditions, and tolerant of head motion, such as functional near-infrared spectroscopy (fNIRS).

fNIRS has gained significant attention in recent years due to its robustness to motion artifacts and environmental noise, making it particularly suitable for applications beyond traditional laboratory settings (Huppert et al., 2006; Pinti et al., 2018; Burgess et al., 2022) and for studying hard-to-test populations (Lloyd-Fox et al., 2010). Similar to fMRI, fNIRS indexes changes in blood oxygenation levels. The increased metabolic demands of highly active brain tissue results in elevated concentrations of oxygenated hemoglobin (HbO) in the blood, as well as reduced concentrations of deoxygenated hemoglobin (HbR), both of which can be measured by shining near-infrared light via a head cap at the wavelengths absorbed by HbO and HbR.

fNIRS provides a cost-effective and versatile alternative to fMRI for studying functional brain activity (Ferrari & Quaresima, 2012). Although it has poorer spatial resolution than fMRI and shallower measurement depth (typically ~3-4cm), it has a higher sampling rate, greater motion tolerance, and fewer practical testing restrictions. Accordingly, fNIRS has attracted growing attention over the last two decades, now with widespread use in developmental research – particularly in infants (Smith et al., 2017; Gu et al., 2017; Weibley et al., 2021; Fiske et al., 2022; Skau et al., 2022; Hancock et al., 2023; Amaireh et al., 2024).

To advance the use of fNIRS in developmental research, it is critical to move beyond demonstrations of feasibility and begin systematically evaluating its reliability and measurement properties across different contexts and timescales – i.e., its neurometric properties (analogous to psychometrics). With regards to fMRI data, test-retest reliability, typically measured via intraclass correlation coefficients (ICCs), is often low, but varies by metric and methodology (Noble et al., 2021; Elliott et al., 2020; Kragel et al., 2021). For example, one meta-analysis found that reliability was particularly low for task-evoked activation at the voxel and regional levels, and for resting-state connectivity at the edge level, but that it improved under certain conditions, such as shorter intervals between sessions and task-based paradigms for estimating functional connectivity (Noble et al., 2021). Relative to fMRI, however, there has been less

methodological research evaluating the reliability and robustness of fNIRS, including its measurement properties (but see Holmes et al., 2024; Yücel et al., 2024).

To determine its utility for developmental research, it is crucial to assess both test-retest reliability across sessions, important for tracking developmental or intervention-related changes, to assess its ability to reliably measure longitudinal changes over development or over the course of an intervention, in addition to its internal consistency (within-sessions reliability), important for studying and understanding individual differences. Relatedly, since reliability can vary depending on measurement conditions and the paradigm employed, it is important to identify the optimal amount of data needed to achieve reliable results for a given paradigm. This complements the growing scrutiny of the variability in collection, preprocessing, and quantification of the fNIRS signals across researchers (Yücel et al., 2025), all of which are needed to enable rigorous scientific investigations using this tool.

While fNIRS has reliably captured group-level effects during cognitive and motor tasks, its ability to produce stable individual-level activation patterns across sessions has not yet been widely examined. Foundational and recent work has begun to address this gap. Plichta et al. (2006) provided one of the earliest demonstrations of fNIRS reliability, showing that visually evoked occipital responses exhibit good test–retest stability over a three-week interval, with high intraclass correlations and reproducible spatial activation patterns. Importantly, they also showed that reliability improves when signals are aggregated across channels, establishing a foundation for later ROI-based approaches in fNIRS research.

Since then, several studies have employed fNIRS to examine test-retest reliability of load-dependent activation in working memory tasks (Yeung et al., 2023; Nittel et al., 2025), motor cortex engagement during finger tapping (de Rond et al., 2023), and resting-state connectivity in clinical populations (Xu et al., 2023). Yeung et al. (2023) reported high test–retest reliability in behavioral performance but only moderate individual-level reliability in neural signals, with regional differences in signal stability. Similarly, while Nittel et al. (2025) reported consistent group-level activation during inhibitory control and working memory tasks, they observed substantial variability in individual-level neural reliability. De Rond et al. (2023) looked more closely at individual-level test-retest conditions in a motor task study, showing high short-term reliability when the fNIRS cap remained in place between sessions, but a significant decline when the cap was removed and refitted 24 hours later. Looking at resting-state fNIRS, Xu et al. (2023) showed strong individual-level test-retest reliability in stroke patients, for global and local efficiency metrics, particularly in low-frequency bands and with longer scan durations. More recently, work from Mollison et al. (2026) further showed that task-evoked responses can exhibit good group-level reliability and that accounting for systemic physiological noise can improve reliability estimates. On the whole, these studies show that group-level test-retest reliability can be high, whereas individual-level reliability tends to be low.

A recent study by Sinfield et al. (2025) examined intra-subject reliability of a single subject across ten repeated fNIRS sessions involving an auditory perception task. This study showed that reliability estimates are highly sensitive to preprocessing choices. Specifically, they showed that accounting for systemic physiological signals increased test–retest reliability, whereas

short-separation channel regression reduced reliability but substantially improved model fit and reduced prediction error, thereby more accurately reflecting task-related neural activity. These findings suggest that highly reliable physiological and extracerebral signals can artificially inflate reliability estimates, highlighting that reliability and validity do not necessarily increase together. Thus, preprocessing decisions can substantially influence both the magnitude and interpretation of reliability metrics. Collectively, these studies establish that fNIRS can yield reproducible signals, but they also highlight that reliability is highly contingent on paradigm, preprocessing, and analytic choices, and remains inconsistently characterized at the individual level.

Building on prior fNIRS reliability work, which has largely focused on low-level sensory paradigms, ROI-averaged activation, or short-term stability under controlled conditions (Plichta et al., 2006; Blasi et al., 2014; Niu et al., 2013; Yeung et al., 2023; Nittel et al., 2025), the present study advances the field by evaluating reliability within the contexts most relevant to developmental cognitive neuroscience: higher-order cognition, individual differences, and longitudinal change under realistic data constraints.

With respect to functional connectivity, even less is known. Although recent studies have used fNIRS to examine inter-regional coherence, network-level dynamics, and functional connectivity (Chenot et al., 2021; Arredonado et al., 2022; Zhilin et al., 2023; Zou et al., 2023), the stability of these metrics within and across sessions remains largely unexplored. These gaps in the literature highlight the need for further methodological advancements to enhance the reproducibility of fNIRS measures before they can be reliably applied in clinical or developmental contexts. Here, we aim to evaluate the robustness of both the within- and between-session reliability of task activation and functional connectivity, both at the level of individual channels and aggregated groups of channels in a region.

A key methodological consideration is how much data is sufficient to ensure robust and interpretable estimates of fNIRS activation and connectivity. While longer task durations can improve signal quality and statistical power by enhancing the signal-to-noise ratio, this comes with diminishing returns; indeed, prolonged data collection could compromise the ability to measure the cognitive processes of interest. Long sessions can result in attentional decline and participant fatigue, on one hand, or practice-related performance improvements on the other. Additionally, prolonged wear of the tight-fitting fNIRS cap can lead to physical discomfort, an aspect that is especially important for developmental populations. Thus, identifying the optimal data quantity needed to achieve stable, high-quality neural metrics is of great importance. This is particularly consequential for developmental and clinical populations, who are more susceptible to cognitive overload and less tolerant of lengthy procedures and long testing sessions. As such, we conducted a systematic evaluation of how activation strength, variability, and statistical significance change with increasing block count, aiming to inform fNIRS study design in sensitive populations, and in time-limited and sensitive research environments.

To address these questions, we collected fNIRS data using a paradigm previously employed in behavioral and fMRI research to investigate relational reasoning. Prior fMRI studies of relational reasoning paradigms consistently implicate two broad regions, lateral prefrontal cortex (LPFC) and posterior parietal cortex (PPC), as central hubs in the reasoning network (Christoff et al.,

2001; 2003; Jung & Haier, 2007; Prado et al., 2011; Krawczyk, 2012; Bunge & Vendetti, 2014; Holyoak & Monti, 2021). Within these broader regions, rostrolateral prefrontal cortex (RLPFC) and inferior parietal lobule (IPL) have been linked to higher-level *relational thinking*—or the capacity to represent, compare, and integrate relationships among sets of stimuli—for example in the service of analogical reasoning (e.g., Christoff et al., 2001; 2003; Bunge et al., 2005; Green et al., 2006; Wendelken et al., 2011; Green et al., 2017). Moreover, reasoning performance has been linked to both lateral frontoparietal functional connectivity and white matter tract coherence (Shokri-Kojori et al., 2012; Cocchi et al., 2013; Ebisch et al., 2013; Bazargani et al., 2014; Wendelken et al., 2017).

While numerous fNIRS studies have studied the neural basis of executive functions (e.g., Cutini et al., 2008; Moriguchi & Hiraki 2013; Pinti et al., 2018), this is, to our knowledge, the first study to use fNIRS to examine relational reasoning. As such, we begin by evaluating whether this neuroimaging method can reliably detect activation and functional connectivity patterns consistent with prior findings from fMRI. To this end, we placed optodes over bilateral LPFC and PPC and measured fNIRS activation and functional connectivity on a variant of the relational matching task previously used in behavioral and fMRI studies involving both adults and/or children (Christoff et al., 2003; Dumontheil et al., 2010; Wendelken et al., 2011; Starr et al., 2022). This paradigm was designed to manipulate relational complexity (Halford et al., 1998) while minimizing other cognitive demands. In the task variant used here, participants assessed similarity among colored shapes, with two levels of relational complexity (see **Figure 2**). 1st-order trials involved direct comparisons of features (color or shape) between two paired images, whereas 2nd-order trials involved comparing two pairs of images, and determining whether they matched along the same feature (color or shape). Thus, successful completion of 2nd-order trials required integration of two relations. Prior fMRI research comparing blocks of trials of each type on relational matching tasks has shown stronger LPFC and PPC activation in response to 2nd-order than 1st-order blocks of trials, particularly in RLPFC and IPL (Christoff et al., 2003), and often more selectively in the left hemisphere (Wendelken & Helskog, 2009; Wendelken et al., 2011).

Our primary fNIRS analyses focused on HbO concentration, given its higher signal-to-noise ratio compared to HbR. We conducted analyses at both the single channel level (38 channels) and aggregated across multiple neighboring channels into 10 regions of interest (ROIs)—bilateral RLPFC, dorsolateral PFC (DLPFC), ventrolateral PFC (VLPFC), IPL, and superior parietal lobule (SPL)—to test whether average ROI metrics would be more stable than individual channels. Given its widespread use in fNIRS research, the ROI aggregation approach was employed alongside all channel-level analyses to evaluate the validity of anatomically defined ROI grouping and to assess the associated trade-offs in spatial specificity and statistical robustness.

A subset of participants completed two sessions spaced 2–4 weeks apart, allowing us to assess the between-session and within-session reliability of behavioral performance, task-evoked activation, and functional connectivity using GLM-based activation estimates and channel-level Pearson and ICCs of HbO concentration. In addition to task-evoked activation, we examined functional connectivity, measured as correlations between fluctuating HbO over time, enabling

us to assess whether network-level coordination of cortical regions during reasoning tasks is consistent over time.

In sum, we first sought to test whether 2nd-order relational reasoning blocks would elicit greater frontoparietal activation and functional connectivity than 1st-order blocks, consistent with increased relational integration demands. Based on prior fMRI findings, we expected the strongest activation for 2nd vs. 1st-order blocks in channels located around RLPFC and IPL, particularly in the left hemisphere. Next, we evaluated the within- and between-session reliability of task-evoked activation and functional connectivity at both the channel and ROI levels. These analyses also served to assess the validity of aggregating channels into anatomically based ROIs, helping to determine whether, or when, such aggregation may enhance signal stability or reduce spatial specificity.

We then examined how activation strength, variability, and statistical power changed as a function of the number of task blocks included in the analyses, so as to inform design considerations for time-constrained studies. Finally, to explore the inclusivity and generalizability of fNIRS for diverse populations, we conducted exploratory analyses building on recent work highlighting the effects of skin tone, hair color, and hair texture on fNIRS data quality (Holmes et al., 2024; Simmons et al., 2025; Yücel et al., 2024). The overarching goal of this study was to evaluate the validity and reliability of fNIRS to investigate relational reasoning, with the goal of enabling its broader application across a variety of populations and paradigms in future research.

2. Methods

2.1 Participants

2.1.1 Full sample

A total of 84 young adults participated in this study; four participants were excluded due to poor task performance, and an additional seven participants were excluded due to low quality fNIRS data (as per criteria described below), resulting in a final sample of 73 participants ($M_{\text{age}} = 21.03$ years, $SD = 2.09$, range = 18-29) (see **Supplementary Table 1** for demographic information). Participants were students enrolled in a Psychology course who were offered course credit for their participation. The total duration of the study was between 60 and 90 minutes. Participants gave written informed consent before participation, and all experimental procedures were approved by the Institutional Review Board at the University of California, Berkeley. All participants had normal (or corrected to normal) vision, no pre-existing neurological or sensorimotor deficits, and no history of self-reported major psychiatric illnesses.

2.1.2 Functional Connectivity Sample

For our initial functional connectivity analysis, we selected a subset of 36 participants with the highest quality data (no more than 8 bad channels out of 46) for pairwise channel-level (channel x channel) connectivity analyses (see **Supplementary Table 1** for demographic information). We additionally performed ROI-level connectivity analyses on both this subsample and the full final sample of 73 participants.

2.1.3 Test-Retest sample

To evaluate the stability of task activation and functional connectivity, we recruited a subset of 26 participants who had high-quality data during session 1 and were willing to return 2-4 weeks (mean = 20.31 days) after their first session (see **Supplementary Table 1** for demographic information). All of these participants met our standard inclusion criteria, and were therefore included in the main analyses. From this test-retest subset of 26 participants, we additionally conducted follow-up functional connectivity analyses on 17 of these participants who had no more than 8 bad channels. ROI-level results for the full sample, reported in the main text, were almost identical to those obtained for the 17 participants with the highest quality data.

2.2 fNIRS Data Acquisition

fNIRS data were collected from a continuous-wave NIRSport2 device (NIRx Medical Technologies, LLC). The wavelengths of emitted light (LED sources) in this system were 760 nm and 850 nm. The data were collected at a sampling rate of 5.1 Hz using the NIRx acquisition software, Aurora fNIRS. The fNIRS cap contained a total of 16 sources and 15 detectors, with the 16th detector used for 8 short channel detectors, creating 38 total channels covering bilateral LPFC (22 channels) and PPC (16 channels). The optode montage (**Figure 2**) was designed using fOLD (fNIRS Optodes' Location Decider; Morais et al., 2018) in combination with an MNI-to-EEG cap converter to optimize coverage of lateral prefrontal and parietal cortices. Guided by prior relational reasoning studies (Wendelken et al., 2008, 2012; Wendelken & Bunge, 2010; Watson & Chatterjee, 2012; Vendetti & Bunge, 2014), target locations were mapped to the international 10–20 system using the Brodmann Brain Atlas. Ten sources and seven detectors were positioned to cover bilateral dorsolateral (BA 9/46), rostrolateral (lateral BA 10), and ventrolateral (BA 45/47) PFC, while six sources and eight detectors targeted bilateral inferior (BA 39) and superior (BA 7) parietal lobules. Analyses were conducted at both the channel and ROI levels, with the 36/38 channels that met our criterion for signal quality being grouped into these 10 anatomically defined regions (see **Figure 1C** and **Supplementary Table 2**).

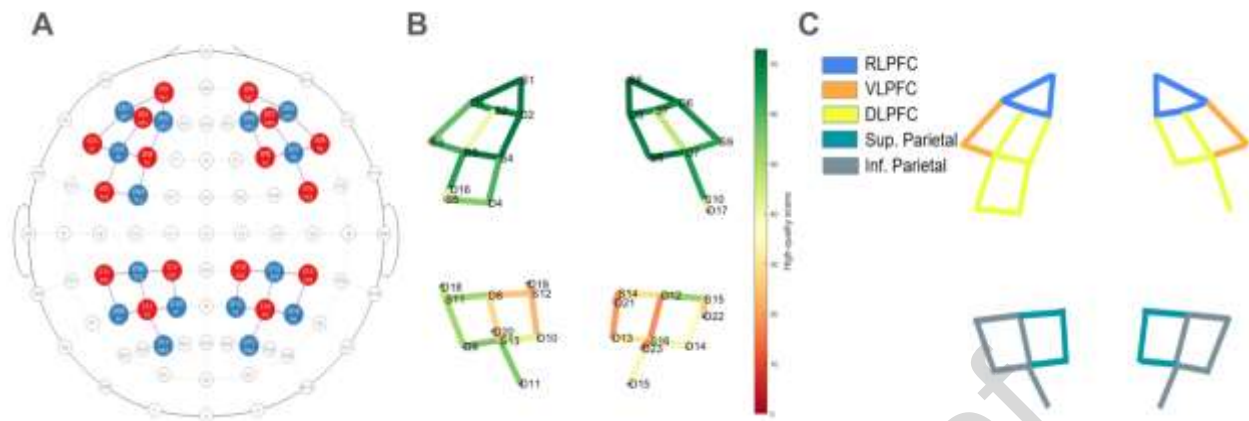


Figure 1. **A:** fNIRS Montage created using the NIRSite software showing the source and detectors in their respective locations that cover bilateral frontal and parietal cortex **B.** Montage figure generated by the QT NIRS package in MATLAB showing the amount of good quality data for each channel in its location within the montage indicating a higher proportion of good quality data in the frontal channels than parietal channels **C.** Colored coded montage diagram showing how the channels that were aggregated into anatomical ROIs.

2.3 Experimental Procedures and Task Design

Participants were prescreened online at the time they signed up for the study. Upon arrival, they completed an additional prescreening survey to ensure they had no neurological, sensory motor, or any self-reported major psychiatric disorders that would exclude them from participating in the study. They then provided informed consent and completed a demographic survey. Once all the intake and pre-screening procedures were completed, experimenters measured the participants' head to determine cap size and placement, then began to set up the cap while participants were given task instructions. The cap was then placed on the participant's head, moving hair as needed to provide clear access to the scalp for the sources and detectors. Cap alignment was verified based on the international 10–20 location of Cz (Klem et al., 1999). fNIRS data were then calibrated and checked for quality using the Aurora fNIRS Software before proceeding with the experiment. If any channels did not have sufficiently high quality signal (see **Section 2.4.2** for criteria), placement and hair-clearing were performed again.

The task was coded and presented using the experiment software PsychoPy (Peirce et al., 2019). Visual stimuli consisted of a stimulus array of four simple colored geometric figures. Two figures were displayed in the top row and two in the bottom row, thereby facilitating the comparison of relations among objects within a row and across rows. Stimulus arrays were presented pseudorandomly and included eight distinct shapes presented in one of eight colors. These figures were combined in unique pairings on each trial to eliminate the possibility of familiarity effects.

The relational matching task consisted of three trial types: Color, Shape, and Match. Color and Shape blocks involved 1st-order judgments, and Match blocks involved 2nd-order judgments (**Figure 2**). On each trial, participants saw two pairs of objects, with a line separating the top pair from the bottom pair. In 1st-order trials, the objects varied either in color or shape. A cue word (“Color” or “Shape”) simultaneously indicated whether participants should focus on the top or bottom row and the type of judgment they should be making. On Shape and Color trials, respectively, participants determined whether the two objects in the cued row had the same shape or color. For example, if the word ‘Shape’ appeared next to the top row of objects, that indicated to the participant to make a shape judgment on the top set of objects. On 2nd-order trials, participants determined whether the relationship between the objects in the top pair was the same as the relationship between the bottom pair of objects. Here, the cue word “Match” was presented beside the line separating the two rows. Across the entire experiment, there were equal numbers of valid and invalid trials for each of the three trial types

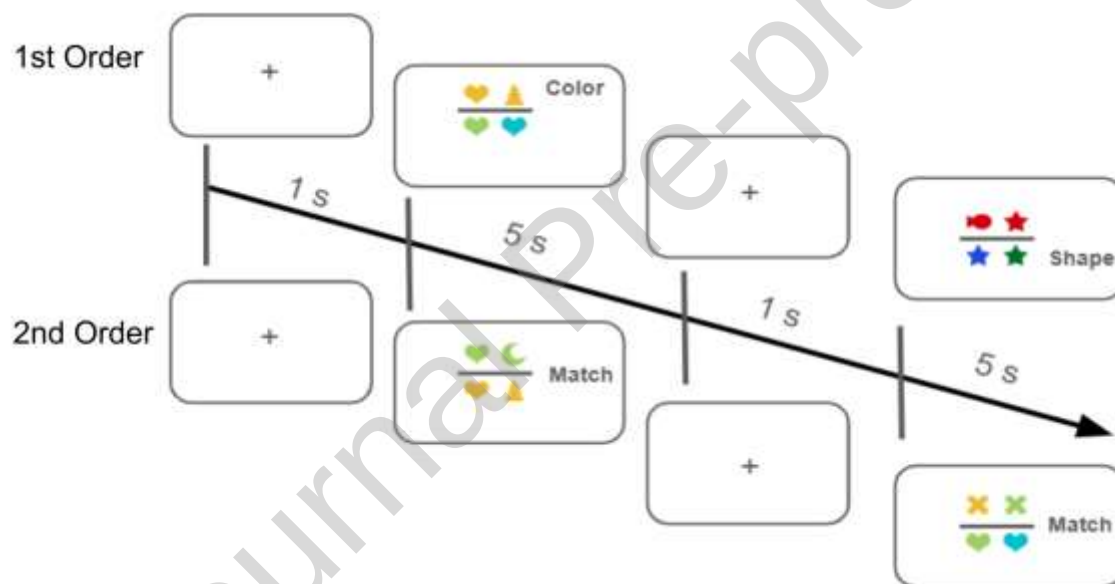


Figure 2. Participants completed a relational matching task consisting of 1st-order (Color and Shape) and 2nd-order (Match) blocks. Each trial displayed two object pairs, with a cue indicating whether to judge shape, color, or relational similarity. Shape and Color blocks required simple feature matching (e.g., “Are these the same shape?”), while Match blocks required comparing the relationship between the top and bottom pairs. Trials lasted 5 seconds with a 1-second inter-trial interval. Blocks consisted of six trials (36 seconds total), presented in a pseudorandomized order and interleaved with jittered rest periods (16–20 seconds). The full session included 9 1st-order blocks, 9 2nd-order blocks, and 19 rest periods (~16.7 minutes total).

The task was presented in a blocked design, with task blocks including 6 1st-order trials (with Shape and Color trials intermixed) or 6 2nd-order trials. Individual trials were 5 seconds long (stimulus remained on the screen for the entire duration regardless of response time (RT)), with

a 1-second inter-trial interval. Individual task blocks were 36 seconds long, and were interleaved with jittered rest periods of 16-20 seconds (16, 18, or 20 seconds, with equal numbers of rest periods of each duration presented pseudorandomly across the experiment). Task blocks were presented in a random order for each participant, and were always followed by a rest period. The entire experiment consisted of a total of 9 1st-order blocks, 9 2nd-order blocks, and 19 rest periods, for a total duration of ~1000 seconds (16.7 minutes). There was a break half-way through, and participants indicated via a button press when they were ready to move on.

Prior to the start of the experiment, participants were given instructions, followed by five untimed practice trials with verbal feedback for each block type. Prior to data acquisition, they performed five timed trials for each block type without feedback, to familiarize them with the pacing of the task. During these practice trials, researchers monitored their understanding and adherence to task instructions.

3. Analyses

All statistical analyses were performed using the pandas and scipy libraries in Python (v. 3.11; Van Rossum & Drake, 2009). Four types of statistical analyses were used to determine the relationship between variables: Pearson's correlations, ICCs, robust regressions, and permutation-based significance testing. The significance level was set to $p < 0.05$, and post-hoc p -values were corrected for multiple comparisons using the Holm-Bonferroni procedure (Abdi, 2010). Results are reported based on group-level statistical parametric maps and false discovery rate (FDR)-corrected q -values ($q < 0.05$).

3.1 Behavioral performance

Accuracy and RTs were quantified by calculating task accuracy for each participant, defined as the proportion of correct responses on each condition across all 18 blocks of trials. Participants whose accuracy was below 70% on one or both conditions (3 participants on 1st-order blocks and one additional participant on 2nd-order blocks) were excluded from further analysis, leaving a subsample of 80 participants whose data were subsequently screened for fNIRS data quality.

3.2 fNIRS quality check and pre-processing pipeline

Because fNIRS signals are sensitive to superficial physiological noise (Santosa et al., 2018), we identified and excluded individual channels and participants exhibiting low-quality data.

Following a visual inspection of each participant's data in the NIRS Toolbox, we conducted a channel-wise quality assessment on the raw fNIRS data using the QT-NIRS module (Montero-Hernandez & Pollonini, 2019), which integrates multiple metrics to identify low-quality measurements. QT-NIRS implements a data-driven algorithm to evaluate signal quality based on physiological and signal characteristics. The criteria outlined below were applied for each channel and participant. Any channels that failed to meet these criteria were flagged as bad channels and removed from subsequent analyses.

Several thresholds were configured: we set the `qThreshold` to 0.7, specifying that a channel must pass at least 70% of quality checks to be retained; the standard `fCut` parameter was set to [0.5-2.5] Hz, specifying the frequency range used to estimate the presence of physiological oscillations (e.g., heartbeat, respiration). This frequency content is extracted using a bandpass filter, and its power is analyzed to assess the presence of expected physiological signals, indicating functional optode-skin coupling. We selected a `sciThreshold` of 0.7, using the Scalp Coupling Index, a correlation-based metric that compares short- and long-separation channel signal shapes to assess optode contact quality. Channels with SCI values below this threshold were classified as poorly coupled. Additionally, the `pspThreshold` evaluation is based on Peak Spectral Power (PSP). PSP calculates the ratio of spectral power in the expected physiological band (typically around the cardiac frequency) to the total signal power. A low PSP indicates a lack of physiological oscillations in the signal, suggesting poor data quality due to insufficient physiological engagement or sensor detachment. By setting a threshold 0.1, channels with a peak spectral power below this level are flagged as low-quality.

We also calculated the number of participants with high-quality data per channel to assess overall signal quality and excluded channels that had insufficient high-quality data across participants from subsequent analyses. An *a priori* power analysis conducted using G*Power indicated that a minimum of 23 high-quality data points per channel was required to achieve 75% power ($1 - \beta = 0.75$) to detect a moderate effect size (Cohen's $d = 0.5$) at an alpha level of 0.05 using a one-tailed test. Based on these criteria, four of the 46 channels, 2 in right SPL and IPL (S14_D13, S16_D12), and 2 short separation channels (S14_D21, S16_D23) were excluded from subsequent analyses (see **Supplementary Figure 1**). Additionally, one channel (S12_D10) and two short channels (S12_19, S13_D20) located over left SPL consistently produced signal quality values that were near or at the exclusion threshold across multiple participants, highlighting a broader issue in fNIRS research: superior parietal cortex is notoriously difficult to image reliably due to increased skull thickness (Beauchamp et al., 2011; Tachtsidis et al., 2016; Skau et al., 2022; Klein, J.H., 2024).

To identify and exclude participants with poor fNIRS data quality, we calculated the mean and standard deviation of the number of bad channels across all scans (mean = 16.12, SD = 8.63). Participants for whom the number of bad channels exceeded 1 SD above the mean (i.e., >27 bad channels) were excluded from further analysis (see **Supplementary Figure 2**). This approach downweights noisy channels in the linear model (see next section). Including these channels does not increase the likelihood of a false positive effect, but the statistical power to detect an effect in this area is reduced (Santosa et al., 2018). Based on this exclusion criterion, 7 of the 80 participants with good behavioral data were identified and excluded from further analyses resulting in a final sample of 73 participants.

After excluding low-quality channels, fNIRS data were analyzed using the NIRS Toolbox in MATLAB. Raw .nirs files containing light intensity measurements were imported and converted to optical density. We applied motion correction using both temporal derivative distribution repair (TDDR) to remove any motion artifacts from the optical density data (OD), and resampled the data to 1.5 Hz. We then converted the cleaned OD data into oxygenated (HbO) and deoxygenated (HbR) hemoglobin concentrations using the modified Beer-Lambert law (Strangman et al., 2003). After each participant's data were cleaned and converted to HbO and HbR concentrations, participant-level statistics were calculated. All subsequent analyses focused on HbO, given its higher signal-to-noise ratio compared to HbR (Pinti et al., 2019). Results for HbR are provided in **Supplementary Figure 4**.

2.4.4 fNIRS Signal Processing

fNIRS data have distinct properties that are not adequately addressed by conventional fMRI-based analysis methods, which can lead to inflated type I error rates (Huppert, 2016). Compared to fMRI, fNIRS data are more strongly affected by serially correlated errors due to its higher sampling rate relative to the underlying physiological signals, as well as by heavy-tailed noise distributions resulting from motion artifacts and substantial variability in signal-to-noise ratio across channels and participants (Huppert, 2016). To address these challenges, a first-level general linear model was applied to each participant's data using an autoregressive, iteratively reweighted least-squares (AR-IRLS) approach. The AR-IRLS model utilizes an autoregressive filter known as 'pre-whitening' to correct for serially correlated errors, along with robust weighted regression to iteratively down-weight motion-related outliers and physiological noise (Barker et al., 2013). This pre-whitening filter is applied to both the participant-level design matrix and the resulting data to remove correlated noise inherent in continuous fNIRS signals. In addition, short-separation channels were included as nuisance regressors in the model to account for superficial physiological signals, thereby further isolating cortical hemodynamic responses of interest and improving the specificity of the estimated neural effects. This model generates participant-level regression coefficients along with their error-covariance matrices, which are used for individual statistical tests, contrast estimation, and subsequent group-level analyses.

In line with current research on the sensitivity and specificity of fNIRS basis sets depending on signal quality and task duration (Santosa et al., 2019), a canonical haemodynamic response function (HRF) basis set was chosen for this analysis. Santosa et al. (2019) found that for task blocks lasting more than 10 seconds (as in this study), the canonical HRF offers optimal performance in a sensitivity-specificity (ROC) analysis. Compared to a full deconvolution of the hemodynamic response (finite impulse response functions, or FIR model), the canonical model has greater degrees of freedom, making it better able to detect a true signal in the presence of noise while improving both sensitivity and specificity in the analysis (Santosa et al., 2019).

Next, group-level statistical models were calculated, using the full covariance matrix from the first-level models to perform a weighted least-squares regression (Santosa et al., 2018) with

condition as a fixed effect and subject as a random intercept ($\beta \sim -1 + \text{cond} + (1|\text{subject})$). Robust regression was also applied to the group-level model to down-weight outliers at the group-level. The results of this analysis were used for the following group-level contrasts for each individual channel: 1st-order activation compared to baseline, 2nd-order activation compared to baseline, the 2nd-order vs. 1st-order contrast, and the Task activation vs. Baseline contrast (1st- and 2nd-order combined into a single 'task' condition). Group activation results are reported as statistical maps using channel-level beta values (mean task-related change in hemoglobin concentration) and t values (beta divided by SE) with corresponding p values.

A slightly different approach was taken to conduct our anatomically informed, ROI-based, analysis. Channel-wise HbO signals were aggregated into ten bilateral cortical regions of interest (ROIs) based on Brodmann Areas (see **Table 1** and **Figure 1C**), spanning RLPFC, DLPFC, and VLPFC, as well as superior and inferior parietal areas. For each ROI, signals from predefined source-detector pairs were averaged to produce a single representative time course. Dummy source-detector mappings were then assigned to ensure compatibility with group-level visualization routines. General Linear Model (GLM) analysis was conducted at the subject level using autoregressive iteratively reweighted least squares (AR-IRLS), which corrects for serial correlations and outliers in the time series as described above. Group-level inference was performed using a linear mixed-effects model with condition as a fixed effect and participant as a random intercept ($\beta \sim -1 + \text{cond} + (1|\text{subject})$), and dummy coding enabled for condition comparisons. A planned contrast comparing condition 2 to condition 1 was computed using a contrast vector $[-1, 1, 0]$, and corresponding statistical maps were visualized and exported. All participant- and group-level results were exported for further statistical analysis.

Benjamini-Hochberg false-discovery rate-corrected p -values (e.g., q -values; Benjamini and Hochberg, 1995) were applied to all group-level analyses, including 36 channels (10 for ROI aggregation analysis), both oxy- and deoxy-hemoglobin signals, and three experimental conditions, making the correction highly conservative across all tests.

Taken together, spatial and physiological noise were addressed through multiple converging safeguards rather than a single correction step: (1) QT-NIRS-based channel and participant exclusion informed by short-long signal relationships, (2) motion correction (TDDR), (3) AR-IRLS modeling to handle serially correlated noise and outliers, and (4) short-separation regression to remove superficial physiology.

3.3 Effects of Number of Task Blocks on Task Activation

To determine the amount of fNIRS data required to obtain stable and interpretable neural activation estimates, we examined t -values and channel-level variance based on incremental block inclusion at both the group and subject levels using activation estimates from the 2nd vs. 1st-order Contrast and the Task vs. Baseline contrast. Specifically, we reanalyzed the data in intervals of 3 blocks (i.e., using the first 3, 6, 9, 12, 15, and all 18 blocks) to evaluate how the quantity of task data influenced the reliability and strength of activation patterns. For each increment of block count, group-level general linear models (GLMs) were computed and beta

and t-values as well as standard error (SE) were extracted to assess how activation evolved as a function of increasing amounts of data. For each block count, we calculated summary statistics including the mean of beta values and SE across channels, as well as the proportion of significant channels. Beta values and SE were first averaged across channels within each participant before being aggregated across participants for each block count.

This two-tiered fNIRS analysis allowed us to quantify data stability as a function of the number of blocks. Importantly, this is not necessarily a case of “more is better”: while increasing the amount of data can enhance statistical power and reduce noise, it may also introduce trade-offs, such as participant fatigue (e.g., attentional drift) and practice-related effects (e.g., increased automaticity or shifts in task strategy). To interpret fNIRS effects across the session in the context of behavioral performance, we also measured accuracy and RT intervals of 3 blocks (i.e., using the first 3, 6, 9, 12, 15, and all 18 blocks). This approach can guide the design of future studies involving young children or clinical populations with limited tolerance for extended scanning, as well as tasks prone to pronounced practice effects, by empirically determining the number of blocks needed to obtain reliable estimates of task-evoked neural activity.

3.4 Functional Connectivity Analyses

Functional connectivity analyses were also performed using the NIRS Toolbox in MATLAB (Santosa et. al, 2018). First, the preprocessed time series for HbO concentration (see above) across the entire ~1000 s collected for each participant was segmented by condition (1st-order and 2nd-order), resulting in a total of 1,500 samples (1.5 Hz × 1000 s) for each participant and condition. Functional connectivity was calculated on HbO only, given its higher signal-to-noise ratio than HbR (Pinti et al., 2019). Functional connectivity was computed using robust pairwise correlations between all channels within each condition block, employing an autoregressive (AR) innovations approach to remove autocorrelation artifacts and isolate inter-regional synchrony and short-separation channels as nuisance regressors to account for superficial physiological signals, thereby further isolating cortical hemodynamic responses of interest and improving the specificity of the estimated neural effects. Connectivity estimates were computed separately for each condition (1st-order, 2nd-order, and Task vs. Baseline), with all data from each block included in the connectivity estimates.

To evaluate condition-level differences in connectivity at the group level, a linear mixed-effects model was fit to the subject-level correlation coefficients (R) using the formula $R \sim -1 + \text{cond} + (1|\text{subject})$, treating condition as a fixed effect and subject as a random intercept. This model enabled estimation of condition-specific connectivity patterns while accounting for within-subject dependence. Results were exported in .csv format for further statistical and visualization analysis in Python. By computing Pearson's correlations between all channels, we obtained a coefficient for each pair – i.e., a pairwise correlation matrix of r values – for each participant for each condition. We then computed the average r coefficient of each element of the matrix across participants, resulting in a global matrix, showing the channel-level connectivity at the group level for channel level (resulting in a 36×36 correlation matrix) and the ROI level (resulting in a 10×10 correlation matrix.) For ROI analyses, we were able to include the full

sample ($N = 73$), excluding channels marked as bad for each participant and averaged over the remaining valid channels within each ROI. For channel-level analyses, we included only a subsample of 36 participants with the highest signal quality, ensuring consistent coverage across source-detector pairs.

For ROI-level analyses, channel-wise time series were aggregated into Brodmann Area (BA)-based regions (see **Figure 1C** and **Table 1**) following the same preprocessing pipeline as the Task vs. Baseline analysis using the NIRS Toolbox in MATLAB. Functional connectivity matrices were estimated using pairwise robust correlations with an autoregressive innovation model (whitening = $4 \times$ sampling rate), and group-level effects were modeled using a linear mixed-effects model with condition as a fixed effect and subject as a random intercept ($r \sim -1 + \text{cond} + (1|\text{subject})$).

To ensure data quality in the full sample ($N = 73$), QT-NIRS-based quality control procedures were applied to exclude low-quality channels on a participant-by-participant basis. Crucially, ROI aggregation enabled full anatomical coverage even in the presence of channel-level signal loss, as each participant contributed usable data to every ROI.

Subsequent functional connectivity analyses were performed using the pandas and scipy libraries in Python (v. 3.11; Van Rossum & Drake, 2009). Three types of statistical analyses were used to determine the relationship between variables: Pearson's correlations, robust regressions, and permutation significance testing. The significance level was set to $p < 0.05$, and post-hoc p -values were corrected for multiple comparisons using the Holm-Bonferroni procedure (Abdi, 2010).

3.5 Within-Session Stability Analyses

To evaluate within-session spatial stability (individual-level) and internal consistency reliability (group-level) of task-evoked activation and functional connectivity, we separated stimulus blocks into interleaved (odd- and even-numbered) blocks and applied the full activation pipeline independently to each subset. For each half, we estimated condition specific task activation using participant-level general linear models (GLMs) and session-level linear mixed effects models (LMMs). Analyses centered on changes in oxyhemoglobin (HbO) signals, due to HbO's greater sensitivity to task-evoked neural dynamics.

3.5.1 Within Session Task Activation

To calculate internal consistency at the individual level, beta values for the Task vs. Baseline and 2-order vs. 1st-order contrasts were extracted for each channel for each participant. ICCs for absolute agreement (ICC(2,1)) were used to quantify internal consistency reliability. We used ICC(2,1) because our central reliability question concerns absolute agreement: whether the same individual shows the same magnitude of response when measured again (across sessions) or estimated from an independent half of the same session. Unlike consistency-based ICCs, ICC(2,1) penalizes systematic shifts in measurement level, making it appropriate when

both rank order and absolute values are of interest. This formulation corresponds to a two-way random-effects model with single-measure reliability, allowing generalization beyond a specific split or session while providing a stringent and interpretable estimate of intraindividual reliability. Cross-classified mixed effect models where halves of the data are nested 1) within-subjects, 2) within-channels, and 3) the interaction between subject and channel ($\beta \sim 1 = (1|\text{subject}) + (1|\text{channel}) + (1|\text{subject:channel})$) to provide an estimate of internal consistency reliability of 1) the fNIRS signal on average at the individual-level, 2) stable channel-level variation in activation, and 3) variation across channels in the degree of reliability at the individual-level. This variation in reliability across channels was visualized by calculating the ICC value within each channel separately, and plotted in **Figure 9**. Again accounting for the reduced amount of data in each split-half, Spearman-Brown prophecy correction was applied to these ICC estimates.

For task activation at the group level, t-values for the Task vs. Baseline and 2nd-order vs. 1st-order contrasts were computed for each channel and ROI then extracted into vectors at the group level. At the group level, Pearson correlations were computed between odd and even halves to quantify the spatial consistency of activation patterns (roughly equal numbers of 1st- and 2nd-order blocks in each half). To account for the reduced amount of data in each split half, we applied the Spearman-Brown prophecy correction formula to the resulting reliability coefficients, yielding adjusted estimates of internal consistency for the full-length data (Spearman, 1910; Brown, 1910). In addition to raw t-value correlations, we assessed the stability of variance estimates (t^2), correlating these across split halves.

3.5.2 Within Session Functional Connectivity

For functional connectivity analyses, connectivity values (Pearson's r) for both contrasts were extracted and matched between odd and even block-derived matrices for each channel or ROI pair. To assess the stability of group-level connectivity patterns, a Pearson correlation was calculated between the vectorized connectivity values from odd and even blocks, yielding a condition-specific reliability coefficient. At the individual level, we assessed the internal consistency reliability of individual functional connectivity profiles by correlating connectivity matrices derived from odd and even blocks for each subject.

This within-session correlation approach for interleaved blocks offers a robust index of the internal consistency of activation and functional connectivity estimates, capturing the reliability of network dynamics within a single session. In addition to this interleaved split-half analysis, we also examined change in fNIRS signal from the 1st half to the 2nd half of the first session to evaluate differences in activation across a testing session (see **Supplementary Section 9**).

3.6 Between-Session Stability Analysis

We assessed the stability of behavioral and fNIRS data across Sessions 1 and 2 by recruiting a subset of 26 participants who agreed to return to be tested a second time, with a 2-4 week

interval between sessions. We examined the stability of accuracy and RTs as well as fNIRS-derived task activation, functional connectivity, and variance at both the group and participant levels.

3.6.1 Between Session Task Activation

To calculate test-retest reliability at the individual level, ICCs for absolute agreement were used that mirrored the split-half calculations, this time with the full data for each session as the two data points per subject. ICC(2,1) was chosen because our central reliability question concerns absolute agreement: whether the same individual shows the same magnitude of response when measured again (across sessions) or estimated from an independent half of the same session. Unlike consistency-based ICCs, ICC(2,1) penalizes systematic shifts in measurement level, making it appropriate when both rank order and absolute values are of interest. This formulation corresponds to a two-way random-effects model with single-measure reliability, allowing generalization beyond a specific split or session while providing a stringent and interpretable estimate of intraindividual reliability. Cross-classified mixed effect models were specified where halves of the data were nested 1) within-subjects, 2) within-channels, and 3) the interaction between subject and channel ($\text{beta} \sim 1 + (1|\text{subject}) + (1|\text{channel}) + (1|\text{subject}:\text{channel})$), and because data are not being split in half, no Spearman-Brown prophecy correction is applied to estimate test-retest reliability. Variation in test-retest reliability across channels was visualized by calculating the ICC value within each channel separately, and plotted in **Figure 11**.

To assess test-retest reliability at the group level, task-related activation was estimated using first-level general linear models (GLMs) for each participant and session, followed by second-level linear mixed effects models (LMMs) for each session. Analyses focused on oxyhemoglobin (HbO) signals, with conditions modeled separately to capture task-specific activation. For each model, t-values were computed per channel and condition, aggregated into group-level vectors, and correlated across sessions to assess the spatial consistency of activation for each condition (1st-order, 2nd-order, 2nd > 1st-order, and Task vs. Baseline). To evaluate variance stability, t-values were squared (t^2) to index signal strength irrespective of direction and correlated across sessions for each channel. This approach quantified the reliability of spatial activation patterns and the reproducibility of signal magnitude across sessions at both the group and individual levels.

3.6.2 Between Session Functional Connectivity

We also evaluated the test-retest reliability of group-level functional connectivity by quantifying the consistency of Channel x Channel and ROI x ROI connectivity estimates across experimental sessions. Functional connectivity matrices derived from session 1 and session 2 were first restricted to channel pairs involving oxygenated hemoglobin (HbO), given its greater sensitivity to task-evoked neural responses. For each task condition represented in both sessions, channel-level connectivity values (Pearson's r) were extracted and paired with identical channel and ROI pairs. To assess stability, a Pearson correlation was computed between the connectivity estimates from session 1 and session 2 for each condition, yielding a condition-specific reliability coefficient. Statistically, this procedure quantifies the linear

relationship between corresponding channels and ROI-to-ROI edges across time points, providing a global metric of reproducibility in functional network structure. Higher correlation values indicate greater consistency of task-evoked connectivity at the group level, while lower values suggest increased session-to-session variability in network recruitment.

At the individual level, this analysis assessed the within-subject reliability of task-related functional connectivity patterns across sessions using a pairwise correlation approach. Channel x Channel and ROI x ROI functional connectivity matrices were derived independently for session 1 and session 2 and filtered to include only oxygenated hemoglobin (HbO) channel pairs, which are more sensitive to task-evoked hemodynamic changes. For each participant, connectivity values from corresponding ROI pairs (defined by source-detector combinations) were aligned across sessions. A Pearson correlation coefficient (r) was then computed between the connectivity strengths observed in session 1 and session 2, providing a metric for evaluating the stability of task-induced brain network organization at the individual level.

4. Results

4.1 Behavioral Results

To confirm participant engagement and assess task performance across conditions, we calculated accuracy and RTs for 1st- and 2nd-order trials. Four participants failed to meet the pre-established performance threshold of 70% accuracy for each condition, and were therefore not included in the final analysis. Participants in the final sample ($N = 73$) performed with accuracy above 90% in both conditions (Mean = 91.2, $SD = .023$ and Mean = 94.27, $SD = .031$) for 1st- and 2nd-order trials, respectively. Unexpectedly, accuracy was 3 percentage points higher for 2nd-order trials ($t(72) = -7.77, p < .0001$); see **Supplementary Figure 3**.

Conversely, participants responded slightly but significantly more quickly on 1st-order trials ($M = 1.56$ s, $SD = 0.31$) than 2nd-order trials ($M = 2.15$ s, $SD = 0.50$), $t(72) = -15.55, p < .0001$, as predicted, suggestive of a small speed-accuracy tradeoff.

The unexpected finding of slightly higher accuracy in the 2nd-order condition, which had higher relational complexity, was small but significant. We hypothesize that it is attributable to higher task-switching demands on the 1st-order trials. Unlike previous studies using similar paradigms, the two 1st-order trial types – shape and color matches – were intermixed within blocks, as opposed to being presented in separate blocks. Moreover, in the current study design, participants were cued, in an unpredictable manner, to respond to one of two rows on every 1st-order trial. In sum, we posit that the increased relational complexity on 2nd-order trials was fairly well balanced by other executive function demands on 1st-order trials, rendering the 2nd- vs. 1st-order comparison a cleaner – albeit more subtle – task manipulation with which to identify activation associated with relational thinking than in prior studies.

4.2 Task-Based Activation

Results are reported based on group-level statistical parametric maps, with significant channel-level and ROI-level results defined as those having false discovery rate (FDR)-corrected q -values $< .05$. Channel- and ROI-level results for both the Task vs Baseline and 2nd $>$ 1st Order contrast presented in **Figure 3** and in **Supplementary Table 3**.

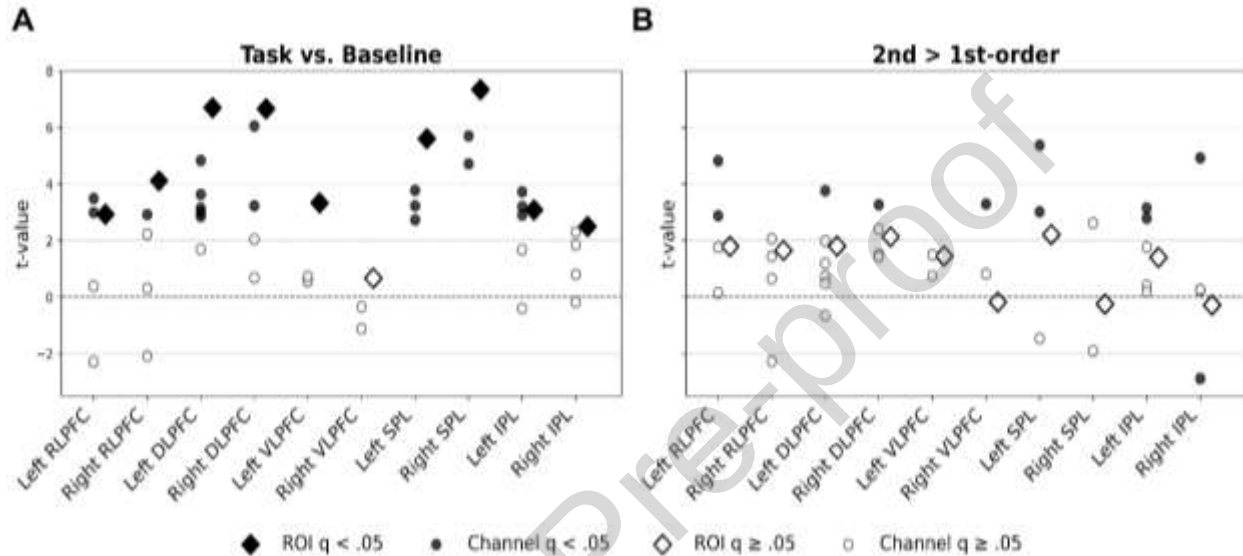


Figure 3. Summary of channel- and ROI-level activation effects. Circles represent channel-level activation and diamonds represent ROI-level activation. Filled symbols indicate effects surviving FDR correction ($q < .05$), whereas open symbols indicate nonsignificant effects. **A.** Task vs. Baseline activation effects, demonstrating widespread significant activation across frontoparietal channels and ROIs. **B.** 2nd-order vs. 1st-order activation effects, demonstrating substantially weaker and less spatially consistent condition-related differences, with only a small number of channels and no ROIs exhibiting significant effects after correction for multiple comparisons.

4.2.1 General Task Activation

To identify cortical regions engaged across both conditions relative to baseline, we performed a contrast comparing neural activation across all 18 blocks (1st and 2nd-order combined). This analysis revealed 20 channels across bilateral prefrontal and parietal cortices that exhibited significantly greater activation for task-related activity (**Figure 3A**; **Figure 4**). Robust activation was observed in the left and right RLPFC and DLPFC channels, with additional activation in a left VLPFC channel, reflecting distributed prefrontal engagement. Significant parietal activation was identified in left SPL and IPL channels, indicating broad posterior involvement. At the ROI level, significant activation increases relative to baseline were observed across nearly all prefrontal and parietal regions (**Supplementary Table 3**) reflecting large-scale cortical recruitment aligned with task demands. While the channel-level analysis provided fine-grained spatial resolution, the ROI-level approach produced stronger and more widespread effects for this general contrast (see **Supplementary Table 3**).

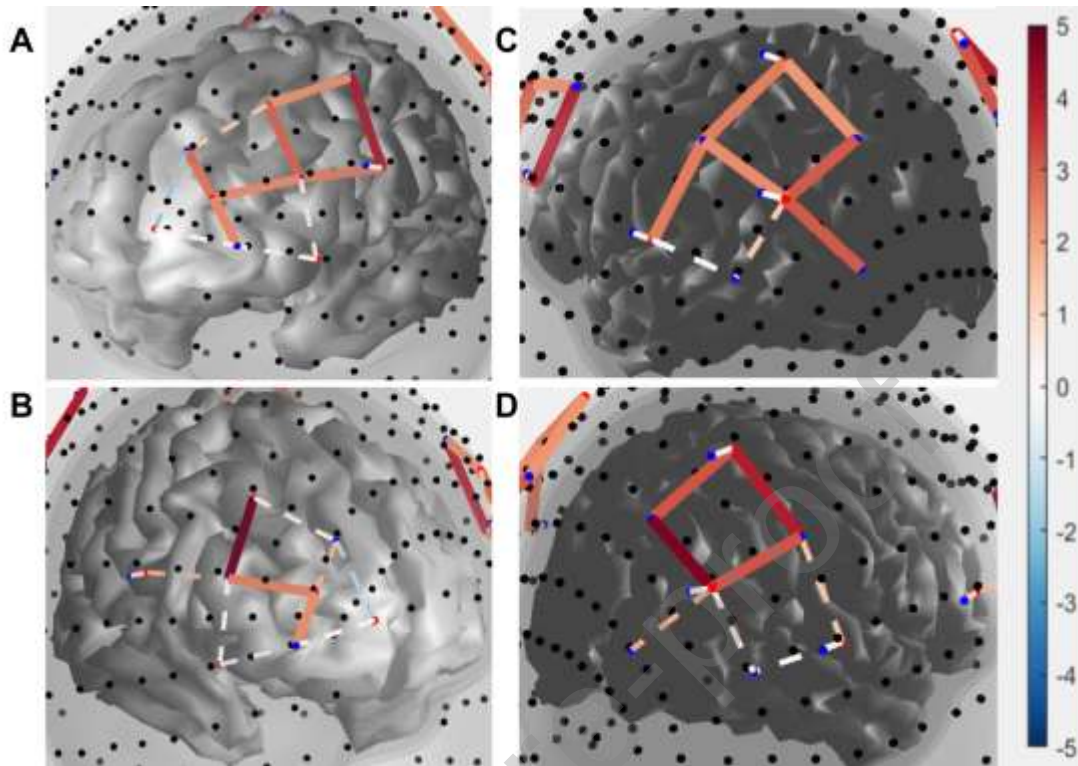


Figure 4. Channel level Task vs. Baseline contrast for task activation across participants. Solids lines indicate channels that were significant after correcting for multiple comparisons across 36 channels; and dashed lines indicate those that did not survive correction. The color bar indicates t-statistic ranging from 5 to -5 **A.** Anterior-lateral view of left PFC showing significant activation in channels over left RLPFC and left DLPFC **B.** Anterior-lateral view of right PFC showing significant activation in channels over right RLPFC and right DLPFC **C.** Posterior-lateral view of the left parietal cortex showing significant activation in channels over left IPL and SPL. **D.** Posterior-lateral view of the right parietal cortex showing significant activation in channels over right IPL and SPL.

4.2.2 Relational Complexity Manipulation

Task-evoked activation at the channel level was significantly greater for 2nd-order relative to 1st-order reasoning in 12 channels distributed across bilateral prefrontal and parietal channels, correcting for multiple comparisons. Robust effects were observed for several channels in bilateral RLPFC and DLPFC, with additional engagement in right VLPFC. Beyond PFC, several channels in left and right SPL and IPL were also significant (see **Figure 3B**; **Figure 5** and **Supplementary Table 3**). By contrast with the channel-level analyses, no ROIs demonstrated statistically significant differences between task conditions after correction for multiple comparisons.

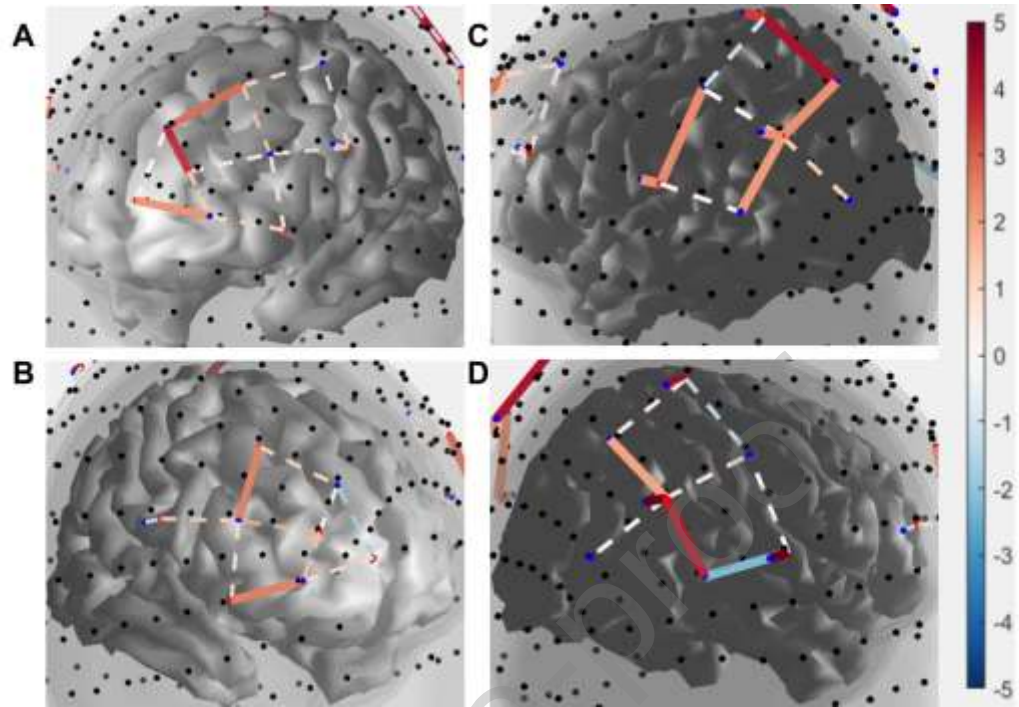


Figure 5. Channel level 2nd vs.1st-order contrast for task activation across participants. Solids lines indicate channels that were significant after correcting for multiple comparisons across 36 channels; dashed lines indicate those that did not survive correction. The color bar indicates t-statistic ranging from 5 to -5. Positive t statistic values indicate greater activation in 2nd Order blocks and negative t statistic values indicate greater activation in 1st Order blocks. **A.** Anterior-lateral view of left PFC showing significantly larger activation during 2nd Order blocks in channels over left RLPFC and left DLPFC. **B.** Anterior-lateral view of right PFC, showing significantly larger activation during 2nd Order blocks in channels over right VLPFC and right DLPFC. **C.** Posterior-lateral view of the left parietal cortex, showing significantly larger activation during 2nd Order blocks in channels over left IPL and SPL. **D.** Posterior-lateral view of the right parietal cortex, showing significantly larger activation during 2nd Order blocks in channels over right SPL and IPL, in addition to a single channel showing significantly larger activation during 1st Order blocks in right IPL.

4.2.3 fNIRS and fMRI Effect Size Comparison

As a final step, we sought to get a rough sense of how robust our fNIRS results were relative to fMRI studies using similar relational matching paradigms (Christoff et al., 2003; Smith et al., 2007; Bunge et al., 2009; Dumontheil et al., 2010; Wendelken et al., 2011a; Wendelken et al., 2011b; see **Table 1**). Of these, we selected the three studies that reported a group analysis of task activation based on an adult sample; all three were based on data collected at 3 Tesla (Christoff et al., 2003; Bunge et al., 2009; Wendelken et al., 2011a).

Focusing on effect sizes reported in RLPFC for 2nd > 1st-order trials, we documented the mean and range of reported – or converted – t-values across channels (for fNIRS) or clusters of voxels (for fMRI) that survived correction for multiple comparisons in the manner specified in these studies. Pooling results from these 3 fMRI studies, t-values for the relational complexity manipulation ranged from 2.62 - 6.65 across the 8 significant clusters in RLPFC. By comparison, t-values for this comparison in the present study ranged from 2.88 - 4.82 across

the 2 channels in LPFC with significantly stronger activation for 2nd vs. 1st-order blocks, correcting for multiple comparisons across all 8 channels. Thus, while there are numerous differences among the studies in terms of sample sizes (with the present one being the largest), measurement granularity (channel-level vs. voxel-level), statistical thresholding approach, areas of LPFC reported, and details of the study design, reported effect sizes between the present fNIRS study and prior fMRI studies are comparable.

Study	Age range	N	RLPFC activations	Mean T-value (min, max)
Christoff et al., 2003	18-25	12	left RLPFC	2.74
Bunge et al., 2009	19-35	15	left & right RLPFC	2.86 (2.62, 3.33)
Wendelken et al., 2011	Young adult	22	left RLPFC	6.65
Current study	18-29	73	left RLPFC	3.85 (2.88, 4.82)

Table 1. Comparison of effect sizes in RLPFC across relational matching neuroimaging studies. N refers to the number of participants. The mean and range of t-values are reported for significant RLPFC clusters (or channels, in the present study). In Christoff et al., 2009, the 2nd > 1st-order comparison corresponds to the no-load internally > externally generated contrast; within an anatomically defined ROI of bilateral RLPFC, significant activation is identified for a single cluster in left RLPFC with an average voxel-level $p < .05$ ($Z > 3.28$), correcting for multiple comparisons. In Bunge et al., 2009, the 2nd > 1st-order task contrast yields clusters in left and right RLPFC that survive correction for multiple comparisons ($p < .05$) across voxels located anterior to the MNI coordinate of $y = +48$. In Wendelken et al., 2011, the 2nd > 1st-order comparison for the visuospatial relational matching task (similar to the present task) identifies a cluster in left RLPFC that survives correction for multiple comparisons ($p < .05$) across the whole brain. In the present study, two RLPFC channels, both in the left hemisphere, survive correction for multiple comparisons across all 8 RLPFC channels ($q < .05$).

4.3 Effects of Cumulative Number of Task Blocks

4.3.1 Task Activation

We visualized channel-level activation patterns at the group level in increments of three cumulative blocks – that is, the first 3, 6, 9, 12, or 15 blocks, or all 18 blocks (**Figure 6**). Additionally, we aggregated over channels to provide numerical comparisons of average channel-level beta values and SE across these increments (**Figure 7**). We also examined how the significance of each individual channel changed status (where significance status was defined as activated, deactivated, or not significant, based on our threshold of $q < .05$) as a function of number blocks for both contrasts (**Supplementary Figures 8 & 9**). In addition to examining the cumulative effect of task blocks included, we conducted complementary analyses

comparing the first and second halves of the session (i.e., independent sets of data (**Supplementary Section 9**).

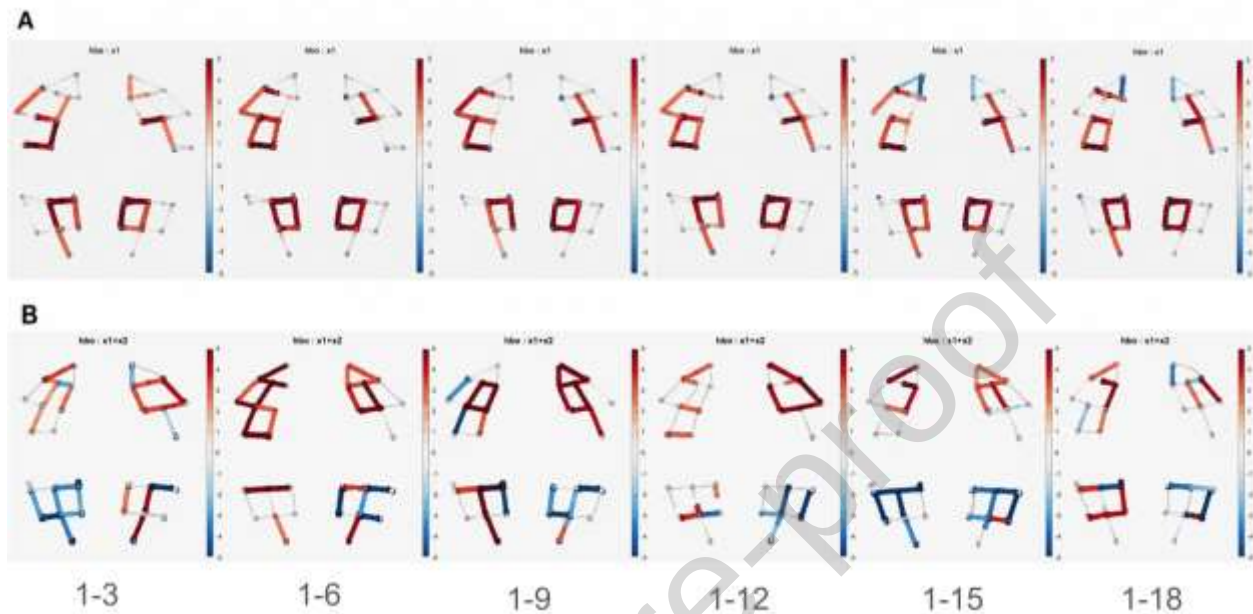


Figure 6. Effects of cumulative block count on group-level activation estimates. Group-level activation maps are shown after cumulatively including the first 1-3, 1-6, 1-9, 1-12, 1-15, and all (1-18) task blocks, showing the results one would have obtained had the task been shorter. Colors represent t-statistics from the group-level mixed-effects model (range = -5 to +5), with warmer colors indicating greater activation and cooler colors indicating reduced activation relative to the comparison condition. **A. Task vs. Baseline** contrast, the overall spatial pattern of activation remains relatively consistent across cumulative block counts, with frontoparietal activation becoming progressively more stable as additional data are included. The greatest changes occur between the earliest block counts, whereas maps generated from 9 or more blocks show comparatively little qualitative change, suggesting convergence of the group-level activation pattern by block 9. **B. 2nd-order vs. 1st-order** contrast exhibits substantially greater variability across cumulative block counts. Both the magnitude and spatial distribution of condition-specific effects change as additional blocks are incorporated, reflecting the lower stability of this more subtle cognitive contrast.

The Task vs. Baseline activation topographic maps suggest a fair degree of stability in the pattern of activation in the aggregate across the 3-block increments (**Figure 6A**); this qualitative observation is supported by relative stability in channel-level significance status (positive, negative, or no activation at $q < .05$; **Supplementary Figure 8**). Quantitatively, the mean group-level beta coefficients and SE both trended downwards across block counts, suggesting improved sensitivity to detect overall task activation with more data, but with diminishing returns after 9 blocks (see **Figure 7A**).

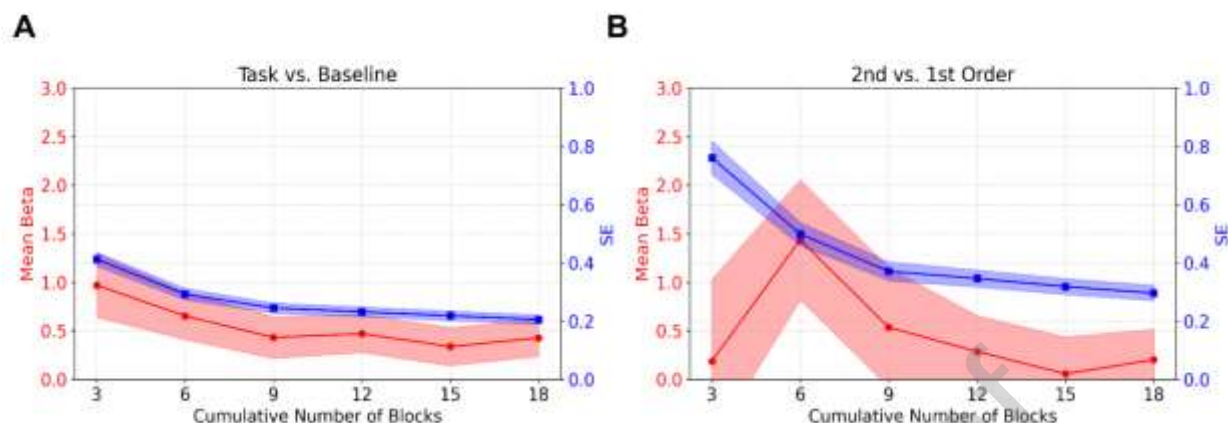


Figure 7. Mean beta values (in red) and SE (in blue) with 95% confidence interval (shaded) across channels as a function of block count, for the first 3, 6, 9, 12, or 15 blocks or all 18 blocks, for **A.** Task vs. Baseline and **B.** the 2nd vs. 1st-order contrast.

For the 2nd vs. 1st-order contrast, the topographic plots hint at a lower degree of stability in the pattern of activation across the 3-block increments (**Figure 6B**); indeed, many channels changed in significance status across the session for this contrast (**Supplementary Figure 9**). On average, mean beta values peaked around block 6 and then decreased, while SE declined steadily across block increments (**Figure 7B**). These aggregated results suggest that statistical estimates for this contrast became more precise as more task data were included, improving precision and a stabilization of condition-specific effect estimates and allowing the model to more accurately distinguish between nuanced differences between conditions. That said, these results must be considered in light of low stability in significance status at the channel level (**Supplementary Figure 9**). Moreover, SE values for the specific contrast were higher at the end of the experiment than for the general contrast; reflecting the general instability of condition level contrasts in neuroimaging research (Heilicher et al., 2022; DeYoung et al., 2025), and/or the possibility that additional blocks of each condition could result in higher stability.

Together, the block count analyses for both contrasts suggest that, while smaller numbers of blocks may produce artificially elevated variability and effect sizes, using at least 9 blocks improved the stability of group-level contrast maps. Behaviorally, too, accuracy and RTs stabilized after approximately 9 blocks (**Supplementary Figure 5**). That said, estimates of task activation became increasingly stable with 15–18 blocks, as indicated by the tapering SE and the convergence of mean beta values. Importantly, however, individual channel-level results reveal marked instability in the 2nd > 1st-order contrast, with substantial flux in significance status across blocks (**Supplementary Figure 8 & 9**).

4.3.2 Behavioral Data

We also examined effects of block count on behavioral performance, with a view to aiding the interpretation of effects on fNIRS activation. To explore changes in task performance across the ~15-minute experiment, given prior evidence of practice-related improvements in task-switching efficiency over time (Karch et al., 2009; Schwarze et al., 2025), we examined performance

across cumulative block counts of 1-3,1-6,1-9,1-12,1-15, and 1-18 (**Supplementary Figure 5**) and in binned block counts (1-3, 4-6, 7-9, 10-12, 13-15, 16-18) (**Supplementary Figure 6**). Accuracy remained consistently high on 2nd-order blocks throughout the session, indicating strong and sustained task engagement. In contrast, accuracy on 1st-order blocks showed modest early variability, which may reflect initial adaptation to task demands, but did not exhibit a clear or consistent pattern of improvement over time. RTs showed only small and variable changes across blocks for both conditions, with no systematic evidence of slowing or performance decline. Overall, these patterns indicate stable behavioral performance across the session, with no evidence of task disengagement.

We observed a high degree of variability in performance, and paired t-tests revealed modest numerical improvements in accuracy that were not statistically reliable for either condition (1st Order: $t(46)=1.57$, $p=.123$, $dz=0.23$; 2nd Order: $t(40)=2.07$, $p=.055$, $dz=0.32$), along with a modest, non-significant numerical speeding of RTs on 1st Order trials ($t(46)=-1.61$, $p=.114$) and no evidence of change on 2nd Order trials ($t(40)=-0.52$, $p=.603$). These results do not provide evidence that would be consistent with instability being driven by a different behavioral pattern in earlier blocks. Early neural instability likely reflects the imprecision of estimating channel-wise GLM effects from small numbers of blocks, particularly for a subtle condition contrast.

4.4 Functional Connectivity

Functional connectivity analyses were conducted at the channel- and ROI-levels for the subset of 36 participants with the highest channel signal quality, as well as on the ROI-level for both this subset (**Figure 8**) and the full sample of 73 (**Figure 9**). As functional connectivity patterns were nearly identical between 1st- and 2nd-order blocks in the full sample (channel level: $r = .997$ ($p < 1.00 \times 10^{-30}$); ROI level: $r = .995$ ($p < 1.00 \times 10^{-30}$), all subsequent connectivity analyses were conducted collapsing across conditions.

At the channel level (**Figure 8**), task-evoked functional connectivity revealed extensive synchronization across bilateral LPFC and PPC; particularly among channels within an ROI, but also between ROIs. At the ROI level (**Figure 9**), the strongest pairwise connectivity for the full sample was found among PFC channels – in particular between left RLPFC and left VLPFC, and between left and right RLPFC – and to a lesser extent within PPC. Both the channel- and ROI-level analyses converged in identifying a coherent frontoparietal network characterized by bilateral and long-range connections. ROI-level connectivity patterns observed in the 36 participants with the highest quality data closely mirrored that of the full sample of 73 (**Supplementary Figure 7**), highlighting the potential to exclude substantially fewer participants when aggregating across channels by dropping bad channels from the ROI estimates. Thus, while channel-level analysis provided finer spatial resolution and revealed specific node-level

interactions, aggregation of channels into ROIs enabled robust connectivity estimation across the full sample while excluding poor-quality channels.



Figure 8. Channel x Channel functional connectivity matrix on the subset of participants with highest quality data (N =36).

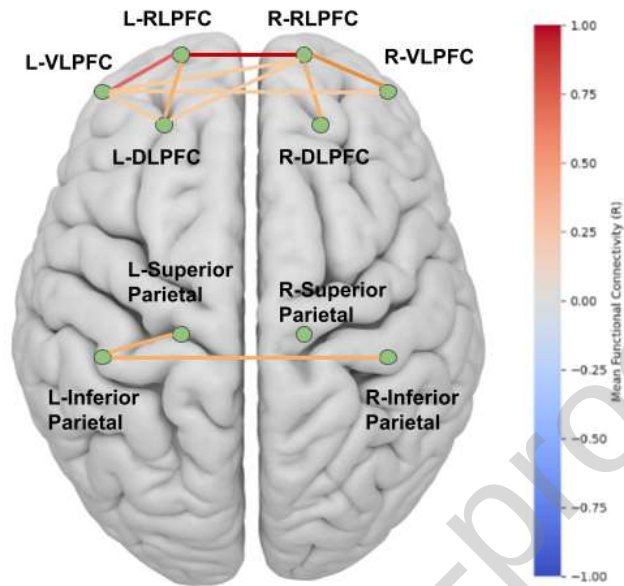


Figure 9. Dorsal view schematic illustrating the strongest patterns of pairwise connectivity ($r > .3$) in ROI x ROI functional connectivity on the full data set ($N = 73$).

4.5 Within-Session Stability

4.5.1 Task Activation

We assessed within-session internal consistency reliability of task-evoked activation within individuals with beta values across interleaved (odd vs. even) blocks, both at the channel and ROI levels. At the channel level, the Task vs. Baseline contrast showed moderate internal consistency reliability on average ($ICC = .54$), with additional appreciable variance attributable to channel ($ICC = .07$) and the interaction between channel and subject ($ICC = .12$). Indeed, there was substantial variation in ICC values across channels, ranging from $ICC = .00$ to $.79$ (**Figure 10A**). On average, individual channels in LPFC ($M_{ICC} = .63$) had significantly higher ICC values than channels in PPC ($M_{ICC} = .49$; $t(34) = 2.84$, $p = .008$). At the ROI level, the Task vs. Baseline contrast showed moderate internal consistency reliability and substantial variance across ROIs in reliability estimates (see **Supplementary Section 10.2**).

By comparison with Task vs. Baseline, the 2nd vs. 1st-order contrast showed lower reliability ($ICC = .44$; **Figure 10B**), with little-to-no variance attributable to channel ($ICC = .00$) or the interaction between channel and subject ($ICC = .00$). When calculating reliability per-channel, there was a wide range of ICC values (see **Supplementary Figure 11**; range = $.00$ to $.54$). On average, individual channels in LPFC ($M_{ICC} = .18$) had similar ICC values to channels in PPC ($M_{ICC} = .24$; $t(34) = -1.16$, $p = .26$). At the ROI level, the 2nd vs. 1st-order contrast again showed

lower reliability, with substantial variance across ROIs in reliability estimates (see **Supplementary Section 10.2**). Overall, the Task vs. Baseline contrast was more reliable than the 2nd vs. 1st-order contrast, and demonstrated moderate within-session reliability across individuals with channel-level metrics consistently outperforming ROI-level estimates; however, these internal consistency estimates varied substantially across channels.

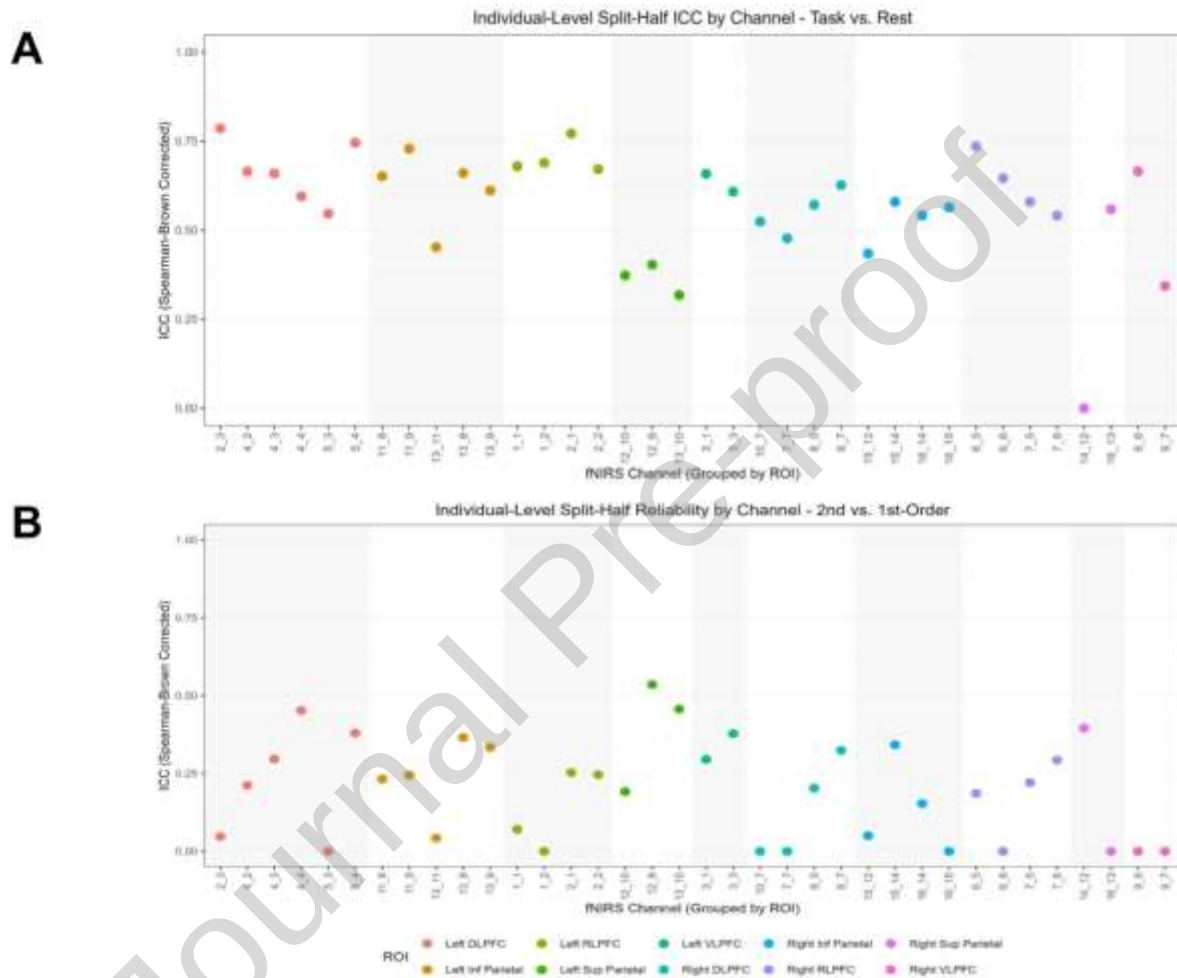


Figure 10. Within-session participant level split half reliability using interleaved blocks for **A.** Task vs. Baseline **B.** 2nd vs. 1st-order contrast. Spearman-Brown corrected ICC values are plotted for each channel, grouped and color-coded by brain region.

We also assessed internal consistency reliability at the group level, for both the channel and ROI spatial scales. Group-level results at the channel level for the Task vs. Baseline contrast showed high within-session spatial stability for interleaved blocks ($r = .60$, $p = .75$, variance = .97 $p < .001$). However, the 2nd vs. 1st-order contrast for interleaved blocks showed low spatial stability ($r = .06$, $p = .12$, variance = .44, $p = .71$) (see **Supplementary Section 10.1**). At the ROI level, Task vs. Baseline demonstrated near perfect internal spatial consistency ($r = .89$, $p = .94$, variance = .87, $p < .001$), while the 2nd vs. 1st-order yielded zero reliability ($r = -.31$, $p = -.91$, variance = .18, $p = .38$; **Supplementary Section 10.2**). Thus, aggregating data across channels

in an ROI and across individuals resulted in very high stability for the general contrast but the specific task contrast was still unstable.

Channel- and ROI- wise ICCs are presented as descriptive estimates to characterize the distribution of reliability across channels, rather than as a series of inferential tests. Accordingly, we do not interpret individual ICCs in a thresholded “significant vs. non-significant” manner, but instead summarize their overall magnitude and variability. Multiple-comparisons correction was applied wherever inferential statistics were conducted (FDR correction for activation maps and Holm–Bonferroni for post-hoc tests), but not to the ICC values themselves, as these are not used for hypothesis testing

4.5.2 Functional Connectivity

To assess the internal consistency of task-based functional connectivity within a session at the individual and group levels, we computed Pearson correlations between connectivity matrices derived from odd and even task blocks, separately for channel-level and ROI-level analyses.

Individual-level within-session stability was extremely high across both spatial scales. For the 36-participant subset, channel-level connectivity correlations between odd and even blocks ranged from $r = 0.944$ to 0.986 (mean = 0.97 , median = 0.97 , SD = 0.01), with all participants exceeding $r = 0.94$. This consistency reflects robust preservation of fine-grained network patterns within individuals over the course of the session (see **Supplementary Section 10.3**).

Group-level analyses also revealed exceptionally high reliability at both spatial scales for both the channel and ROI-levels (see **Supplementary Section 10.3**). At the channel level, the correlation between connectivity matrices across all 630 unique channel pairs was $r = 0.998$ ($p = 2.36 \times 10^{-4}$, permutation test), indicating near-identical network structures between odd and even blocks. Thus, functional connectivity showed very high stability within a session.

4.6 Between-Session Stability

4.6.1 Behavioral Data

To confirm participant engagement and assess task performance across conditions, we examined accuracy and RTs for 1st-order and 2nd-order trials across sessions 1 and 2. All participants exceeded the pre-established performance threshold of 70% accuracy in both sessions, and were therefore included in the final analysis. As in session 1 (1st-order = 91.2%, 2nd-order = 94.3%), accuracy in session 2 was near ceiling (1st-order = 91.27%, 2nd-order = 95.36%), and differences between the sessions were marginal ($t(n-1) = -2.02$, $p = .054$), with a trend toward improvement in Session 2 and indicating stable performance from session 1 to 2. Due to limited inter-individual variability in accuracy, we did not conduct a correlation analysis to assess task performance stability. Conversely, RTs consistently favored the 1st-order condition, with faster responses in both Session 1 (1.53 s vs. 2.02 s; SEs = 0.0345, 0.0559) and Session 2 (1.40 s vs. 1.71 s; SEs = 0.0345, 0.0559). Paired-sample t-tests revealed significantly faster RTs in Session 2 than Session 1 ($t(n-1) = 5.16$, $p < .001$). RTs showed strong within-subject

stability across sessions, with strong correlations for both 1st-order ($r = .75$, $p < .001$) and 2nd-order ($r = .86$, $p < .001$) conditions. (see **Figure 11**). For full statistics see **Supplementary Section 11.1**.

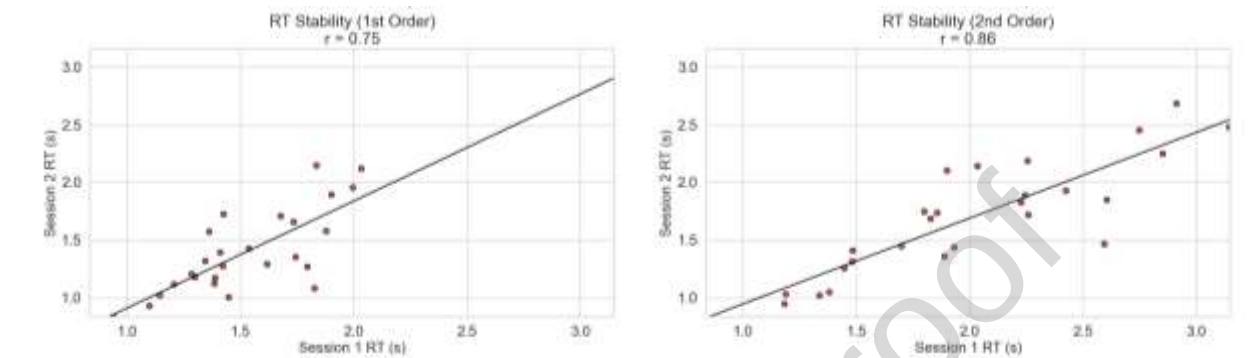


Figure 11. Between session RT stability for **A.** 1st-order and **B.** 2nd-order conditions. Individual data points correspond to average RTs for each individual at session 2 plotted against session 1.

4.6.2 Task Activation

Individual-Level

To assess the test-retest reliability of task-evoked neural responses across sessions, we examined the stability of beta values across sessions at the channel and ROI levels within participants. At the channel level, the Task vs. Baseline contrast showed relatively low test-retest reliability within participants ($ICC = .18$), with appreciable variance attributable to channel ($ICC = .07$) and to the interaction between channel and subject ($ICC = .05$). Despite the low overall average test-retest reliability, a wide range was observed, $ICCs = .00 - .65$, and there were several channels with good test-retest reliability estimates (see **Figure 12A**). On average, individual channels in LPFC ($M_{ICC} = .35$) showed significantly higher test-retest reliability than channels in PPC ($M_{ICC} = .21$) ($t(34) = 2.23$, $p = .03$). Like the channel-level results for Task vs. Baseline, the 2nd vs. 1st-order contrast showed low test-retest reliability ($ICC = .13$), with very little-to-no variance attributable to channel ($ICC = .00$) or the channel by subject interaction ($ICC = .00$). When examining differences across channels in reliability, over half of the channels showed no reliability, but there were some exceptions, $ICCs = .00 - .45$ (see **Figure 12B**). Aggregating at the ROI level yielded similar results: both contrasts showed relatively low test-retest reliability, but there were notable differences across ROIs in these reliability estimates (see **Supplementary Section 11.2.1**).

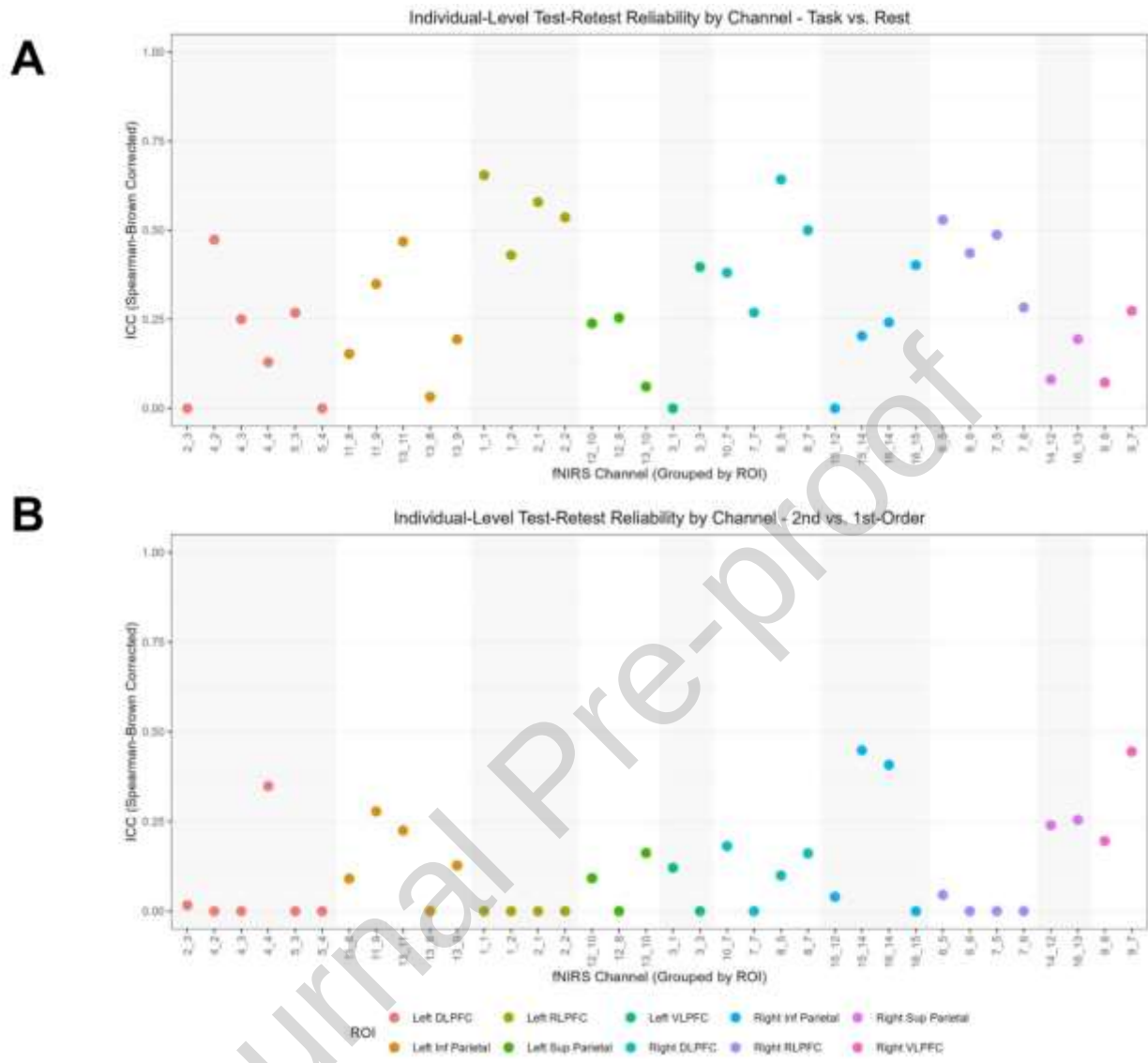


Figure 12. Between-session participant level stability using ICC(2,1) for **A.** Task vs. Baseline contrast indicating moderate reliability for channels in bilateral RLPFC and right DLPFC, and **B.** 2nd vs. 1st-order contrast, indicating low reliability for most channels

Overall, both contrasts demonstrated low test-retest reliability across both spatial scales at the participant level, with channel-level metrics consistently outperforming ROI-level estimates. Task vs. Baseline contrasts showed higher internal consistency reliability than the 2nd vs. 1st-order contrast at the ROI level, but the opposite was true at the channel level. While ROI aggregation improves anatomical interpretability, it may reduce sensitivity to stable individual differences, especially for condition-specific contrasts, underscoring the utility of fine-grained channel-wise analysis in reliability-focused fNIRS studies.

As in the individual level within-session results, Channel- and ROI- wise ICCs are reported as descriptive indices to characterize the distribution of test-retest reliability across channels,

rather than as a set of inferential tests. Accordingly, we do not interpret individual ICC values in a thresholded “significant vs. non-significant” manner, but instead summarize their overall magnitude and variability. Multiple-comparisons correction was applied for all inferential analyses (FDR correction for activation maps and Holm–Bonferroni for post-hoc tests), but not to the ICC values themselves, as these are not used as hypothesis-testing endpoints.

Group-Level

Although test-retest reliability was low at the individual participant level, we sought to assess test–retest reliability of the group-level activation pattern. To this end, we examined both the spatial consistency of t-values and the stability of activation variance (t^2) across sessions and the difference in amplitude from session to session, both at the channel and ROI levels. At the channel level, spatial activation patterns were moderately stable for Task vs. Baseline, but not for the 2nd vs. 1st-order contrast (**Figure 13**). Variance stability had the opposite pattern, with nonsignificant correlations for Task vs. Baseline and high 2nd vs. 1st-order contrast variance stability. At the ROI level (**Supplementary Section 11.2.1**), the Task vs. Baseline contrast showed even greater test–retest reliability ($r = .87$, variance = .93, $p < .001$), while the 2nd vs. 1st-order contrast remained unstable ($r = .28$, variance = .31, $p = .44$) (**Supplementary Figure 13**).

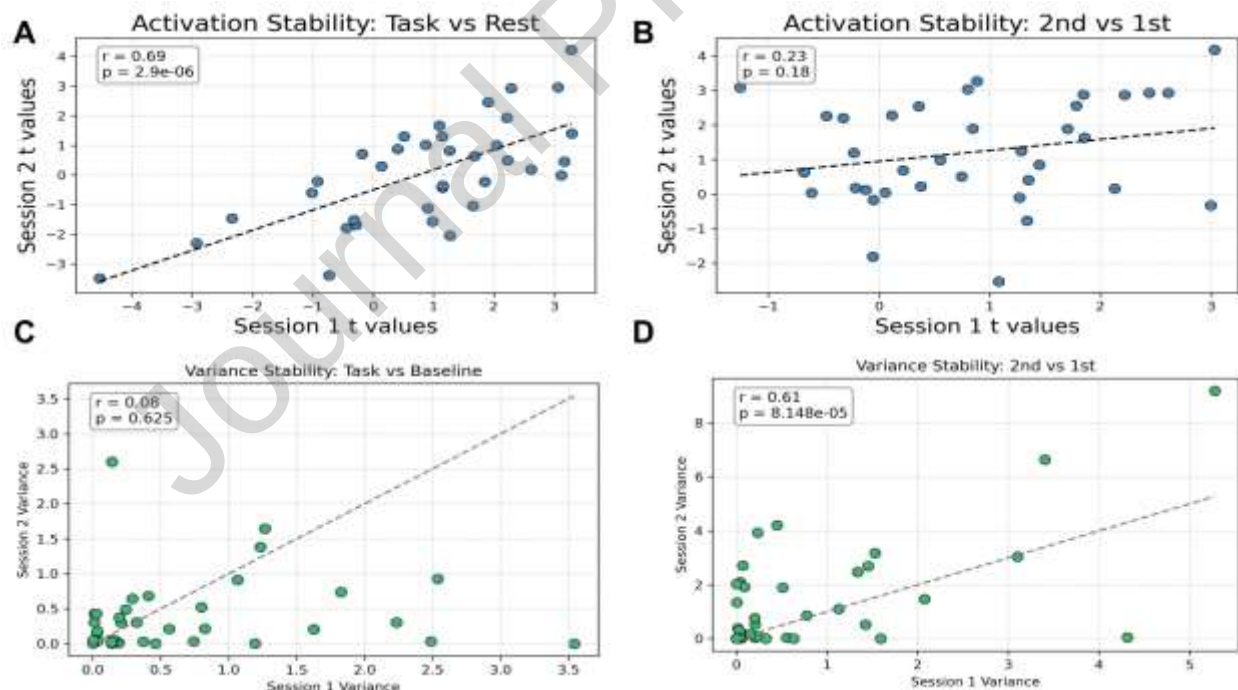


Figure 13. Group-level activation and variance stability data at the channel level. **A.** Task vs. Baseline correlation across sessions indicating strong correlations between sessions ($r = .69$) **B.** 2nd vs. 1st correlation across sessions, indicating very low correlation between sessions ($r = .23$) **C.** Task vs. Baseline variance correlation across sessions, indicating low correlation between sessions ($r = .08$) **D.** 2nd vs. 1st variance correlation across sessions, indicating moderate correlation between sessions ($r = .61$)

With regards to overall signal amplitude, we observed higher channel-level amplitude in session 1 than session 2 for the Task vs. Baseline contrast ($t = 4.03$, $p < .001$). The 2nd vs. 1st-order contrast showed the opposite pattern; however, these results were not significant ($t = -1.28$, $p = .21$). At the ROI level, the pattern of signal amplitude differed across sessions for both the Task vs. Baseline contrast ($t = 3.96$, $p = .004$) and the 2nd vs. 1st-order contrast ($t = 4.90$, $p = .001$), indicating higher ROI level activity in session 1 than session 2. In sum, both group-level spatial and variance-based metrics indicate that ROI-level activation estimates demonstrate greater test–retest reliability than channel-level data for overall task activation; both were poor for the specific task contrast. While channel-level analyses may offer higher spatial resolution and sensitivity to localized effects, they appear to be less stable across sessions at the group level, especially for fine-grained condition contrasts.

4.6.3 Functional Connectivity Stability

To assess the test–retest reliability of individual functional connectivity, we computed Pearson correlations between each participant's session 1 and session 2 connectivity matrices separately at the channel and ROI levels. At the channel level, subject-wise correlations across the 26 participants showed high consistency across all conditions. For 1st-order, correlations ranged from $r = .17$ to $.95$ (mean $\pm$ SD: $.73 \pm .12$); for 2nd-order, from $r = 0.18$ to 0.94 ($.74 \pm .11$); and for Task vs. Baseline, from $r = 0.23$ to 0.95 ($.70 \pm .12$). All correlations were statistically significant ($p < 0.001$, permutation test), indicating robust individual-level stability of channel-wise functional connectivity (**Figure 14B**). At the ROI level, functional connectivity demonstrated even greater stability (**Supplementary Section 11.3.2**). To summarize, ROI-level connectivity exhibited slightly higher test–retest reliability than channel-level data, with tighter correlation ranges and greater central tendency. This pattern confirms the robustness of ROI-level connectivity as a stable and interpretable marker of between-session functional architecture. Both approaches, however, confirmed strong within-subject stability, supporting their utility for longitudinal designs and individualized network characterization.

Group-level test-retest reliability for functional connectivity at the channel level was also extremely strong ($r = 0.97$, $p < 1 \times 10^{-66}$) (**Figure 14A; Supplementary Section 11.3.1**). At the ROI level, similarly high stability was observed (**Supplementary Section 11.3.2 and Supplementary Figure 14**).

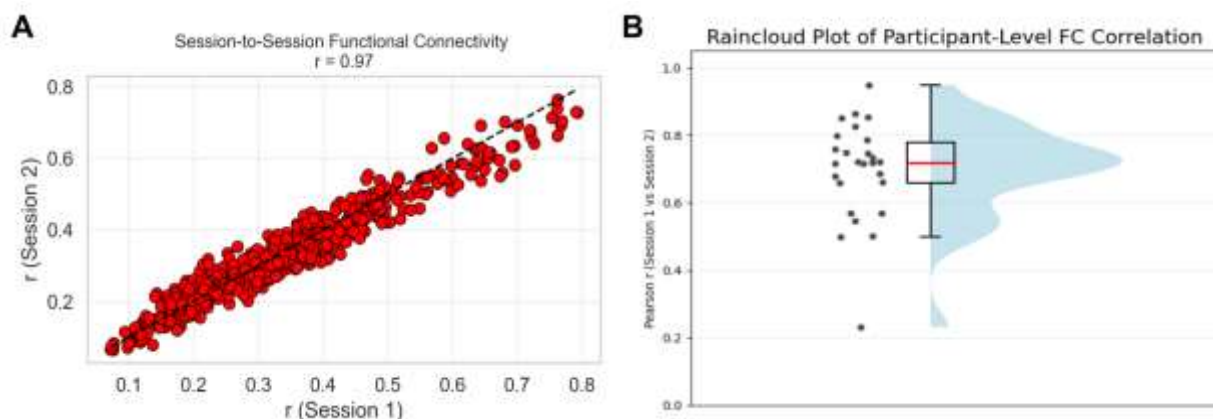


Figure 14. Channelwise functional connectivity correlations between sessions at **A.** Group-level ($r = .97$, $p = 2.28 \times 10^{-66}$) and **B.** Participant level functional connectivity correlations between sessions, ranging from .23 to .95 (mean = .7)

4.7 Exploratory Analysis: Effects of Hair and Skin Types

An exploratory analysis was conducted to assess whether phenotypic characteristics — specifically skin tone, hair color, hair type, and hair texture—predicted fNIRS signal quality, as documented in prior studies (Holmes et al., 2024; Simmons et al., 2025; Yücel et al., 2024). For full methods see **Supplementary Section 12**.

Skin tone significantly predicted bad channel count, $F(3, 50) = 4.07$, $p = .012$. Post hoc Tukey's HSD tests revealed that participants with brown skin had more bad channels than those with fair skin (mean difference = 10.92, $p = .082$), and participants with medium skin showed a similar trend (mean difference = 6.01, $p = .084$), though both effects were marginally significant. Hair color was also a significant predictor of signal loss, $F(2, 50) = 6.71$, $p = .003$. Participants with red or black hair had more bad channels than those with blonde hair (mean difference = 10.56, $p = .007$) and brown hair (mean difference = 5.31, $p = .027$), while the difference between brown and blonde hair was not significant ($p = .297$). In contrast, neither hair type ($F = 0.93$, $p = .43$) nor hair texture ($F = 0.045$, $p = .64$) significantly predicted the number of bad channels (**Supplementary Figure 15**).

Together, these findings suggest that certain phenotypic features—particularly darker skin tones and non-blond hair colors—are associated with greater fNIRS signal loss. Although limited by small and uneven subgroup sizes, this pattern supports prior research on melanin- and hair-related light attenuation and emphasizes the need for inclusive methodological practices and better-powered future studies examining phenotypic bias in optical neuroimaging.

5. Discussion

In the current study, we sought to evaluate the reliability of fNIRS data on a sample of healthy adults during a reasoning task, with a view toward extending it to pediatric populations. While fNIRS technology has great potential for advancing developmental cognitive neuroscience due to its unique advantages of robustness to motion artifacts and environmental noise, making it particularly suitable for applications beyond traditional laboratory settings (Huppert et al., 2006; Pinti et al., 2018) and for studying hard-to-test populations (Lloyd-Fox et al., 2010), the signal quality and neurometric properties of the signal need to be comprehensively evaluated. After using fNIRS to corroborate results from similar fMRI paradigms, we found that signal quality and reliability (both within-session and across-session) varied across different quantification techniques and spatial scales.

The instability observed in condition level contrasts (2nd vs. 1st-order), especially at the individual level, is largely attributed to their computation as "difference scores". This analytical approach can introduce a high degree of measurement error and consequently result in lower

reliability estimates. These same analytical and methodological concerns have been shown in previous fMRI research, resulting in similarly low reliability for contrast-based ROI work. Previous research has suggested that using condition specific beta (β) coefficients, which directly characterize percent signal change, yields greater reliability values, suggesting that previous reports of low fMRI test-retest reliability may stem from methodological considerations in data analysis rather than an intrinsic limitation of the modality itself (Heilicher et al., 2022; DeYoung et al., 2025).

Given known anatomical and physiological challenges in optode-scalp coupling over posterior regions (Beauchamp et al., 2011; Tachtsidis et al., 2016; Skau et al., 2022; Klein, J.H., 2024), particular attention was given to data quality in PPC. Four channels located over right SPL (S14_D13, S14_D21, S16_D12, S16_D23) were excluded due to their poor signal quality and three channels located over left SPL consistently yielded signal quality values near or at the exclusion threshold across multiple participants. Additionally, PPC showed lower reliability in both within and between session metrics compared to LPFC. This reflects a broader issue in fNIRS research: PPC is notoriously difficult to image reliably, due to factors such as greater hair density, increased curvature of the posterior scalp, and a thicker section of skull, all of which attenuates near-infrared light transmission. These factors contribute to higher rates of signal dropout and greater variability in signal-to-noise ratio in parietal areas compared to more anterior regions like PFC. As a result, although these superior parietal channels were retained in the final montage to preserve anatomical coverage and network completeness, their inclusion warrants caution when interpreting activation or connectivity findings involving this region. The challenges underscore the need for region-specific quality checks in fNIRS preprocessing pipelines and reinforce the growing consensus that some cortical areas remain technically constrained for optical neuroimaging.

5.1 Task-Based Activation

Task performance reliably engaged widespread frontoparietal regions, with the Task vs. Baseline contrast revealing significant activation in 20 of 36 channels spanning bilateral PFC and PPC: most notably bilateral RLPFC and DLFC. When aggregating across channels, significant activation was observed in all ROIs: bilateral RLPFC, DLPFC, VLPFC, and parietal cortices, underscoring the value of ROI-level analysis in detecting broad patterns of engagement by increasing statistical power and reducing noise via spatial averaging.

The more selective 2nd vs. 1st-order contrast elicited activation in a subset of 12 channels, predominantly in the left hemisphere. These activations were qualitatively similar to those of prior fMRI studies of relational reasoning, particularly the robust activation observed in left RLPFC (Smith et al., 2007; Bunge et al., 2009; Dumontheil et al., 2010; Wendelken et al. 2011; Wendelken et al., 2011; Christoff et al., 2020; see Vendetti & Bunge, 2014). Moreover, the t-values observed in left RLPFC in our fNIRS data (2.88–4.82) were similar in magnitude to those reported in comparable fMRI studies of relational matching, suggesting that fNIRS provides sufficient sensitivity to detect meaningful differences in relational complexity at the level of individual channels.

In contrast, ROI-level analysis for this contrast did not reveal any statistically significant effects after correction. This highlights a key limitation of spatial averaging: subtle differences between closely matched task conditions may be obscured, particularly when both involve high cognitive demands, as in the current paradigm, where 1st-order shape and color trials were intermixed within blocks, increasing task-switching requirements.

Consistent with prior fMRI reliability work, we note that contrast-based ROI measures often exhibit lower test–retest reliability than estimates based on single-condition betas. Meta-analytic and empirical studies (Elliott et al., 2020; Noble et al., 2021) have shown that univariate task-fMRI contrasts can yield relatively poor ICCs, in part because contrast scores compound measurement error from multiple conditions. Situating our findings within this broader literature helps clarify why more nuanced contrasts in the present study show reduced stability relative to general task effects.

Together, the results for the general and specific task contrasts illustrate a critical trade-off between spatial specificity and statistical robustness across the two approaches to fNIRS analysis. Channel-level analysis excels at detecting localized, condition-sensitive activation, making it especially valuable for fine-grained contrasts. ROI-level analysis, by contrast, enhances sensitivity to broad activation patterns, as shown in the Task vs. Baseline contrast, but may fail to detect more subtle condition differences when spatial averaging reduces functional resolution. The choice between these approaches should therefore be guided by the contrast of interest and the expected spatial distribution of effects.

5.2 Effects of Number of Blocks on Task Activation

Here, we sought to determine how many task blocks are needed to obtain reliable behavioral and neural activation estimates in our paradigm, starting with an adult sample. Our goal was to balance statistical power with participant experience and experimental validity. To this end, we conducted blockwise convergence analysis across the full 18-block session, examining average beta values, variability, and standard error for the general and specific contrasts.

For the Task vs. Baseline contrast, beta values and standard error declined steadily with each additional set of the three task blocks, while the proportion of significant channels increased. These trends suggest improved signal-to-noise ratio with more data, although gains diminished after approximately 9 blocks. While extending the session beyond this point yielded marginal benefits, the most substantial improvements in model precision and group-level sensitivity for the general contrast were achieved within the first half of the experiment.

For the 2nd vs. 1st-order contrast, early cumulative blocks yielded inflated beta values and a high proportion of significant channels, followed by stabilization of both activation magnitude and variability by approximately block 9. The gradual decline in standard error across blocks further supports improved estimation precision with increasing data. Behavioral stabilization during an initial learning phase may also contribute to the observed neural stabilization, even though in this study there was no reliable or systematic early-session learning effect observed at

the group level: accuracy changes were small and variable, and RTs showed only modest, non-significant speeding. With respect to fNIRS signal, then, early instability is expected when estimating channel-wise GLM effects from a small number of blocks, where standard errors are larger, beta estimates are more volatile, and significance patterns are inherently unstable. In this sense, the cumulative analyses primarily reflect a stability/precision tradeoff, whereby reliable estimation, especially for subtle condition contrasts, requires sufficient data.

At the same time, cumulative analyses inherently conflate increasing data quantity with time-in-session, making it difficult to isolate whether observed changes reflect improved estimation or true neural dynamics. Complementary within-session analyses comparing the first and second halves of the experiment therefore provide critical additional insight. These analyses revealed substantial reorganization in the spatial distribution of activation patterns, resulting in low correspondence between halves. Importantly, this pattern would not be expected if the cumulative effects were driven solely by improved statistical precision, suggesting that genuine within-session neural change contributes to the observed dynamics. In addition to these spatial differences, we also observed an overall reduction in activation magnitude across the session, with neural signal amplitude consistently lower in later blocks, indicating that neural responses become attenuated with continued task exposure, potentially as a result of increased neural efficiency and/or task disengagement.

Taken together, these findings suggest that the observed patterns reflect a combination of improved estimation precision with increasing neural and behavioral data and changes in neural signal magnitude. This distinction has important implications for experimental design. Rather than assuming that longer sessions uniformly improve data quality, it is critical to identify a window in which neural signals are both stable and representative of the cognitive processes of interest. In the present paradigm, converging evidence from neural and behavioral data suggests diminishing returns after approximately 9 blocks, marking a potential “sweet spot” for reliable and interpretable contrast estimation. At the same time, the continued decline in standard error in beta estimates up to block 18 suggests that additional data may further improve precision, particularly for subtle contrasts such as 2nd > 1st-order trials, although this remains to be empirically tested.

For this paradigm and population, then, the first nine blocks may provide a practical balance between estimation precision and minimizing the influence of within-session neural adaptation. However, because cumulative analyses confound increasing data quantity with time-in-session, these findings should not be interpreted as establishing a universally optimal task duration. Rather, they highlight the need to consider both statistical precision and potential changes in behavioral and neural responses over the course of a session when designing fNIRS paradigms. In light of these results in this experimental paradigm, a more useful recommendation is that researchers first conduct sufficient behavioral piloting and practice prior to recording to ensure that performance is as stable as possible, minimizing early learning effects and late task disengagement, and then evaluate the quantity of data required to obtain stable neural estimates. Future work should more directly dissociate data quantity from time-dependent neural change, for example through non-sequential sampling or adaptive designs, to further refine estimates of optimal task duration.

This initial exploration therefore provides a methodological foundation for identifying optimal task duration in fNIRS studies, capturing stable behavioral performance, well-separated condition-specific neural signals, and improved group-level sensitivity without incurring the potential costs of longer sessions. Our approach also reinforces the value of combining amplitude-, pattern-, and variance-based reliability metrics, as well as visualizing spatial activation patterns, to fully characterize the temporal dynamics of fNIRS signals in cognitive tasks. However, these analyses will need to be extended to developmental and clinical populations, where sustained attention is limited and long sessions may not be feasible. Optimizing this balance between data quantity and neural stability will be critical for interpreting developmental change and individual differences in future research.

5.3 Functional Connectivity

To evaluate large-scale cortical integration during task performance, we compared functional connectivity analyses at both the channel and ROI levels. No significant differences in connectivity emerged between the 1st and 2nd-order conditions (consistent with unpublished results from our laboratory), suggesting that while relational complexity modulates localized activation patterns, the broader functional integration of the frontoparietal network remained stable across conditions. Accordingly, we collapsed across conditions for subsequent connectivity analyses. Significant connectivity was observed across multiple channels, but the strongest and most reliable pairwise connections were concentrated between LPFC channels and between PPC regions. Notably, left RLPFC channels exhibited the strongest pairwise connectivity, both with left VLPFC and with right RLPFC, consistent with its proposed role as a hub for executive control and relational reasoning.

The ROI-level analysis revealed a functionally coherent and anatomically interpretable frontoparietal network, with significant connectivity both within and across hemispheres. Robust frontal lobe coordination was observed between left RLPFC, DLPFC, and VLPFC, as well as between homologous regions in the right hemisphere. Bilateral frontoparietal integration was also prominent, including strong cross-hemispheric connections between RLPFC and inferior parietal cortex. Notably, left RLPFC again emerged as the most strongly connected ROI, consistent with findings from prior fMRI research involving this paradigm in a pediatric sample (Wendelken et al., J Neuro, 2017).

Crucially, ROI-level analyses produced consistent results across both the full sample ($N = 73$) and a particularly high-quality subsample ($n = 36$), indicating that this approach enables inclusion of more participants in functional connectivity analyses by averaging over variable channel-level quality within each region. This methodological strength preserved statistical power and enhanced generalizability—an important consideration for population-level research or clinical studies in which optode coverage may vary.

Together, these findings underscore the complementary strengths of each analytic approach. Channel-level analysis is highly sensitive to focal, optode-specific interactions and can reveal nuanced connectivity patterns that may be masked by data aggregation. Nevertheless, this approach is also more susceptible to local measurement noise and variability in optode

positioning across participants, potentially leading to higher rates of data exclusion and reduced generalizability. Future research may benefit from combining both analytic levels, thereby leveraging their respective advantages while mitigating individual limitations.

5.4 Within Session Reliability of Task Activation

To evaluate the internal reliability of fNIRS task-based measures within a session, we compared activation and connectivity estimates across interleaved blocks, at both the channel and ROI levels. Interleaved designs offered the most direct assessment of measurement stability while minimizing confounds such as learning or fatigue, serving as a complementary validation of temporal consistency to the comparison of 1st and 2nd halves of the session (**Supplementary Section 9**).

The current findings provide a nuanced view of within-session stability in fNIRS-derived task activation, underscoring the importance of contrast type, spatial resolution, and region of interest. Task vs. Baseline activation demonstrated the highest within-session reliability across all analyses.

At the individual level, the Task vs. Baseline contrast showed moderate internal consistency at the channel level, with notable variability across individual channels and channel-by-subject interactions. LPFC channels exhibited significantly greater reliability than PPC channels, indicating that prefrontal regions may provide more robust signal estimates for broader task activation. This pattern was mirrored at the ROI level, which also showed moderate reliability with clear variability across regions. In contrast, the 2nd vs. 1st-order contrast again showed slightly lower internal consistency at both levels, with minimal variance attributable to specific channels or channel-by-subject interactions. LPFC and PPC regions did not differ significantly in this case, and channel-level ICCs were generally lower. These findings reinforce the conclusion that task complexity and cognitive contrast type play a critical role in determining the stability of measured activation.

At the group level, channel-level spatial patterns were highly consistent across interleaved comparisons, with strong correlation and variance preservation. In contrast, the 2nd vs. 1st-order contrast yielded low spatial stability at the group level. This suggests that this particular condition-specific contrast may not be suitable for investigating the spatial distribution of activation at either the channel or ROI level.

While our activation based internal consistency analyses yield similar results to previous research for both group and participant data at channel- and ROI-levels, our functional connectivity measures showed exceptional within-session reliability, with near perfect split-half correlations at the group level and across individuals. These findings far exceed the reliability levels reported in previous fNIRS studies using activation-based ICCs or overlap metrics. Our results suggest that connectivity-based approaches offer superior reproducibility and may be less affected by spatial noise or individual differences in activation topography.

Together, these findings highlight the critical role of spatial resolution and split-half methodology in assessing reliability. While ROI aggregation provides interpretable group-level summaries for broad contrasts like Task vs. Baseline, it tends to mask finer-grained variability that is captured more effectively at the channel level. Channel-wise analysis proved more sensitive to individual and condition-level effects, particularly for detecting subtle differences in relational reasoning contrasts.

Overall, these results support the value of multi-scale approaches to within-session reliability assessment in fNIRS. Our specific task contrast was highly unstable, whereas the general contrast of task activation versus baseline was more robust. Functional connectivity measures emerged as the most robust indicator of stable brain network organization, demonstrating exceptional consistency across both spatial scales and participant-level analyses. This suggests that connectivity-based metrics may offer a more reproducible alternative to activation-based approaches, especially for longitudinal designs and individual differences research.

5.5 Between-Session Stability of Task Activation and Functional Connectivity

To evaluate the suitability of fNIRS-derived measures for longitudinal and individual differences research, we assessed the inter-session stability of task-evoked activation and functional connectivity at both the channel and region-of-interest (ROI) levels. Sessions were spaced 14–33 days apart, offering a stringent test of short-term stability that provides a ceiling estimate for the reliability expected in developmental or intervention-based designs. By comparing these spatial scales across group- and subject-level analyses, we provide a comprehensive characterization of their test–retest reliability, including not only the consistency of activation patterns but also the stability of signal variance—an often-overlooked dimension critical to interpreting inter-individual variability and signal robustness.

5.5.1 Task Activation Stability

These findings offer a detailed characterization of test–retest reliability in task-evoked fNIRS activation across sessions, highlighting key differences in stability as a function of spatial scale, contrast type, and anatomical region. At the participant level, between-session test–retest reliability was generally low, regardless of spatial scale or contrast. For the Task vs. Baseline contrast, reliability estimates were modest overall, though some individual channels, particularly in lateral prefrontal regions, showed acceptable-to-high levels of stability. ROI-level aggregation, while beneficial for group-level interpretability, did not enhance individual-level reliability – and in some cases reduced sensitivity to consistent individual differences. For the 2nd vs. 1st-order contrast, reliability was again low across the board, with few channels demonstrating meaningful stability and parietal regions slightly outperforming frontal regions. These findings underscore the challenge of capturing stable individual-level activation patterns across sessions, particularly for subtle contrasts.

At the group level, neural signal amplitude was consistently higher in session 1 than in session 2 across both contrasts and spatial scales, suggesting a potential decline in signal magnitude over time. Spatial activation patterns for the Task vs. Baseline contrast were reliably reproduced

across sessions, particularly at the ROI level, highlighting the benefits of spatial averaging for enhancing reproducibility. In contrast, the 2nd vs. 1st-order contrast exhibited poor spatial reliability at both scales, indicating greater vulnerability to session-to-session variability. Variance-based analyses revealed a complementary pattern: signal variance was more stable for the 2nd vs. 1st-order contrast. This dissociation suggests that while spatial patterns of broad task engagement are stable, response magnitude may be more sensitive to task complexity and individual differences.

These results underscore the importance of aligning spatial scale with analytic goals when designing fNIRS studies. At the group level, ROI-based analyses demonstrated clear advantages for detecting stable spatial activation patterns, particularly for broad task contrasts like Task vs. Baseline, due to the statistical benefits of spatial averaging. ROI aggregation mitigates the influence of noisy or missing channels and enhances reproducibility, making it well-suited for studies prioritizing anatomical interpretability and robust detection of task-related activation across sessions. However, this benefit did not extend to our condition-specific contrast, which remained unstable across sessions even at the ROI level. Moreover, variance-based measures revealed that signal variability over time was better preserved for the more specific contrast at the ROI level, despite poor spatial reproducibility.

In contrast, channel-level analyses, while offering finer spatial resolution, showed lower between-session stability on average, particularly at the group level and for condition-specific contrasts. Nevertheless, some individual channels, especially in prefrontal regions, showed acceptable or good reliability, and channel-wise metrics preserved variability in signal amplitude that may be obscured by regional averaging. This suggests that channel-level data may be more sensitive for capturing participant-specific features and should be considered in studies focused on individual differences, longitudinal change, or personalized applications such as neurofeedback or intervention tracking.

Together, these findings support a complementary analytic strategy: using ROI-level aggregation to enhance group-level robustness and interpretability, while retaining channel-level analyses to preserve sensitivity to individual variability and spatially localized effects.

Beyond these methodological considerations, an important feature of these findings is that the observed between-session neural changes occur in the context of relatively stable behavioral performance. At the group level, we did not observe evidence of systematic behavioral decline or disengagement across the session, suggesting that the reduction in neural signal amplitude from Session 1 to Session 2 is unlikely to reflect fatigue or loss of task engagement. Instead, this dissociation—stable performance alongside changing neural activation—points to a scenario in which participants are performing the task equally well while relying on less metabolically demanding neural processes over time. A large body of work across fMRI and fNIRS demonstrates that neural activation, particularly in frontoparietal systems, often decreases with repeated task exposure while performance remains stable, reflecting more efficient allocation of cognitive resources (Haier et al., 1992; Kelly & Garavan, 2005; Hempel et al., 2004; Ayaz et al., 2012; Hernandez et al., 2019; Mandrick et al., 2017). In this sense, neural variability across sessions may reflect changes in how the task is implemented neurally, rather

than a degradation of task performance or measurement quality. This interpretation also has important implications for test–retest reliability: if neural representations shift across sessions while behavior remains stable, reliability estimates based on activation magnitude or spatial patterns may be attenuated. It is therefore plausible that more extensive practice prior to the first session could stabilize neural responses and, in turn, improve between-session reliability. This pattern highlights a key challenge for longitudinal fNIRS research: reliability estimates may be constrained not only by measurement noise, but also by meaningful, experience-dependent changes in neural organization.

5.5.2 Functional Connectivity Stability

In contrast to activation analyses, functional connectivity measures, both at the channel and ROI level, demonstrated exceptionally high test–retest reliability, especially at the group level. Channel-level connectivity matrices yielded near-perfect correlations across sessions for all conditions, indicating the global architecture of functional networks is highly preserved over time, even at a fine-grained spatial resolution. ROI-based matrices showed slightly higher correlations, suggesting a marginal advantage in signal stability, likely due to enhanced signal-to-noise ratio via anatomical aggregation. Importantly, these patterns held across task conditions and in the Task vs. Baseline comparison, indicating that core network topology remains stable across both cognitive state and time.

At the individual level, both spatial scales again produced strong test–retest reliability. Channel-wise connectivity correlations were high for all participants across conditions, with all participants exhibiting significant stability. ROI-level correlations were even stronger with 94% of participants showing correlations exceeding $r = .80$. This pattern suggests that while channel-level functional connectivity captures individualized neural signatures, ROI-level connectivity provides a more robust and generalizable index of inter-session reliability, potentially making it the preferred approach for clinical, developmental, or biomarker research where individual tracking is critical.

As a sensitivity analysis for functional connectivity, we examined a higher-SNR subset in which we included only participants with relatively few bad channels (17/26) to evaluate whether connectivity reliability depended on uniformly strong channel quality. Notably, ROI-level connectivity results in this higher-quality subset were almost identical to those in the full sample, demonstrating that our connectivity findings are not an artifact of selectively retaining only pristine data.

Importantly, because connectivity was estimated using pairwise correlations across full trial epochs, variance measures were not computed separately. However, the consistently strong test–retest correlations at both scales implicitly reflect stability in both connectivity patterns and underlying signal structure, supporting their use as reliable longitudinal markers.

The current study demonstrates higher between-session reliability than previously reported in fNIRS test–retest studies. For example, Yeung et al. (2023) reported only moderate ICCs (0.37–0.62) for task-evoked frontal activation during a working memory task, with limited individual-level reliability for differential contrasts. Similarly, Nittel et al. (2025) found replicable group-level patterns in working memory and inhibitory control tasks but low individual ICCs (< 0.50) and inconsistent overlap metrics. de Rond et al. (2023) showed that activation reliability in motor

cortex declined sharply when the fNIRS cap was removed and refitted, underscoring the impact of optode placement. Xu et al. (2023) demonstrated strong resting-state fNIRS test–retest reliability for graph metrics in stroke patients, particularly when scans were sufficiently long and focused on low-frequency bands.

Our findings extend this literature by incorporating both t-value and signal variance correlations at the group level, as well as ICC and variance attributed to both channel and ROI at the participant level. While the participant level data showed similar results to previous work (ICCs ranging from .00 - .65 for Task vs. Baseline and .00 - .45 for 2nd vs. 1st-order), the group-level ROI-based metrics in our study yielded stronger reliability than those reported previously (e.g., $r = .87$ and variance = .93 for Task vs. Baseline), while functional connectivity far outperformed typical activation-based estimates (channel-level $r > .97$, ROI-level $r > .98$). It is very likely that participant level ICCs for both spatial scales were significantly impacted by the overall poorer data quality in parietal channels compared to PFC.

These results provide a robust foundation for future longitudinal research, highlighting the strengths of combining both ROI and channel-level approaches to capture reproducible, interpretable, and spatially sensitive neural measures. It is important to note that these results reflect only short-term stability (2–4 weeks), which is appropriate in the context of developmental populations where neural and behavioral measures may fluctuate over time. This short-term window provides a practical upper bound, or "ceiling," on the maximum stability that can be expected.

5.6 Within- and Between Session Comparison

A direct comparison of within- and between-session reliability highlights several important trends regarding the robustness of fNIRS-derived activation measures. Across both temporal scales, the Task vs. Baseline contrast demonstrated consistently higher reliability than the 2nd vs. 1st-order contrast, supporting prior findings that condition-specific contrasts—particularly those calculated as difference scores (e.g., Condition 1 - Condition 2)—result in a high degree of error making them inherently less stable and more vulnerable to cognitive variability and noise (Heilicher et al., 2022). This instability may be further exacerbated by practice or strategy-related effects across sessions, which can alter the strength of differential contrasts such as 2nd vs. 1st-order comparisons.

At the channel level, both within- and between-session results revealed high variability in reliability across channels, yet a subset of channels—particularly those located in lateral prefrontal cortex (LPFC)—showed moderate to high reliability in both contexts. These findings suggest that while channel-level data can be noisy and sensitive to signal quality fluctuations, certain individual channels can nonetheless provide robust and reproducible indices of activation, particularly in well-covered and low-noise regions such as the PFC. In contrast, parietal channels consistently showed lower reliability across both within- and between-session assessments, likely due to poorer signal quality and higher rates of noisy or borderline-usable channels.

At the ROI level, overall reliability was lower than at the channel level for individual subjects, especially for the 2nd vs. 1st-order contrast. However, some regions, namely bilateral rostralateral prefrontal cortex (RLPFC), demonstrated relatively greater stability across sessions. Importantly, the direction and magnitude of reliability estimates for specific ROIs sometimes differed between within- and between-session comparisons, suggesting that aggregation may obscure meaningful subject-level variance or amplify inconsistencies across time. This highlights the need to interpret ROI-based results with caution, particularly when assessing individual-level effects in longitudinal designs.

Functional connectivity analyses further revealed exceptional stability across both within- and between-session assessments. At the individual level, channel-wise connectivity showed robust reproducibility across participants, and ROI-level measures yielded even higher consistency, with the majority of participants exceeding reliability thresholds commonly required for clinical or biomarker research. At the group level, split-half correlations for connectivity matrices were near perfect across all conditions and both spatial scales, far surpassing traditional activation-based reliability metrics. These findings indicate that the global structure of functional brain networks is highly preserved across time and cognitive states. These results suggest that functional connectivity, especially when derived from ROI-level aggregation, provides a highly stable and generalizable neural signature that is well-suited for longitudinal, developmental, and individualized research applications.

Together, these results emphasize the complementary value of multi-scale analysis in assessing the stability of task-evoked activation. Channel-level analyses may be more vulnerable to variability, but they still capture reliable subject-specific patterns, especially in the PFC. As a result, they are particularly useful for research focused on individual differences. ROI-level analyses, though more robust at the group level, may underperform in capturing reliable participant-level effects, especially for condition-specific contrasts. These findings underscore the importance of aligning spatial scale, contrast type, and analytic goals when interpreting fNIRS data in both cross-sectional and longitudinal research contexts.

These findings offer a cautious but optimistic assessment of fNIRS as a tool for longitudinal and developmental research. At the group level, ROI-based activation analyses of broad contrasts such as Task vs. Baseline yielded high spatial stability ($r > .85$), while functional connectivity measures demonstrated near-ceiling reproducibility ($r > .95$). At the individual level, task activation ICCs were modest at best, often falling below the thresholds required for individual-level inference. However, select prefrontal channels showed acceptable reliability, and individual functional connectivity profiles remained stable across sessions at both the channel and ROI levels. Overall, the most stable data emerged from ROI-level aggregation at the group level and from selective high-quality channels or connectivity matrices at the individual level. A hybrid approach—leveraging ROI-level metrics for group robustness and interpretability, and channel-level or connectivity-based metrics for individual variability—appears best suited for fNIRS applications in clinical, developmental, and longitudinal research contexts.

5.7 Advancing fNIRS Reliability Research

The present study extends prior fNIRS reliability research (Plichta et al., 2006; Niu et al., 2013; Blasi et al., 2014; de Rond et al., 2023; Yeung et al., 2023; Nittel et al., 2025; Sinfield et al., 2025; Mollison et al., 2026) by evaluating measurement stability within the contexts most relevant to developmental cognitive neuroscience—higher-order cognition, individual differences, and longitudinal stability—rather than the tightly controlled sensory paradigms and short-term intervals that have characterized much of the existing literature.

Recent work has increasingly emphasized that reliability estimates are shaped by methodological and analytic choices, and also that reliability should not be equated with measurement validity (Sinfield et al., 2025). In light of these concerns, we employed a conservative preprocessing pipeline designed to minimize physiological and extracerebral contamination and applied identical preprocessing, quality-control, and analysis procedures across all datasets, thereby reducing the likelihood that observed reliability differences were driven by methodological artifacts rather than underlying neural signal characteristics. Building on this emerging perspective, the present study evaluates reliability in a complex reasoning task across a large sample of participants and multiple levels of analysis, including channels, anatomically defined ROIs, and functional connectivity networks. By examining both reproducibility and data requirements in a higher-order cognitive paradigm, our findings extend existing reliability frameworks beyond low-level sensory tasks and provide a more comprehensive characterization of measurement stability in cognitive neuroscience applications of fNIRS.

A central contribution of the current study is the conceptualization of reliability as a multilevel psychometric property. By jointly assessing within-session internal consistency and between-session test–retest stability, we demonstrate that reliability varies systematically across analytic targets, with general task activation exhibiting substantially greater stability than more specific condition contrasts. This provides a more precise and design-relevant characterization of measurement robustness than prior single-metric approaches.

We further extend existing frameworks by incorporating variance as a complementary dimension of psychometric inquiry. In addition to rank-order consistency, we evaluate the stability of effect magnitude and dispersion across channels, ROIs, and sessions, allowing us to distinguish noise-driven instability from meaningful changes in variability—an essential consideration for developmental research focused on individual differences.

Methodologically, we move beyond threshold-dependent and correlation-based metrics by introducing continuous, pattern-based measures of reproducibility, including vector correlations of activation and connectivity profiles. These approaches preserve graded spatial information and provide a more sensitive estimate of distributed neural pattern stability. Importantly, we also demonstrate that task-evoked functional connectivity exhibits high reliability over multi-week intervals, extending prior work largely limited to short-term resting-state assessments and supporting its use in longitudinal and biomarker-oriented applications.

Finally, by quantifying how reliability changes with increasing data, we address a key design question for developmental research, showing diminishing returns beyond approximately seven minutes of data collection. At the same time, our findings demonstrate that apparent improvements in stability can reflect both increased estimation precision and genuine changes in neural responses across the session, underscoring the need to consider reliability within the broader context of task dynamics and neural adaptation.

Taken together, these findings establish a modern, psychometrically grounded framework for evaluating fNIRS reliability, demonstrating that contemporary approaches can robustly capture higher-order cognitive processes and functional network organization. This work therefore provides both conceptual and practical guidance for the use of fNIRS in longitudinal and individual-differences research in developmental populations.

5.8 Impact of Hair and Skin Characteristics

Our findings provide further evidence that individual differences in hair color and skin tone meaningfully impact fNIRS signal quality, underscoring the critical need to incorporate phenotypic characteristics into methodological reporting and study design (cf. Bradford et al., 2024). Specifically, darker skin tones (more melanin) and darker hair colors, particularly black and red (more eumelanin or pheomelanin, respectively), were associated with a significantly greater number of bad channels identified during preprocessing. These effects were observed independently, without significant moderation by hair texture or type, suggesting that melanin-related optical properties and hair pigmentation remain primary sources of signal degradation in near-infrared spectroscopy.

Although not all pairwise comparisons reached conventional significance thresholds, the overall pattern aligns with previous studies demonstrating reduced optode-scalp coupling and increased signal attenuation in participants with darker pigmentation or denser hair coverage (Holmes et al., 2024; Simmons et al., 2025; Yücel et al., 2024). The present findings reinforce the argument that hair and skin phenotypes should not be treated as nuisance variables or omitted from consideration, but rather as legitimate sources of biological variance that can systematically bias data quality.

These results call for standardized inclusion of skin tone and hair phenotype metrics in all future fNIRS reports, particularly in the context of sample description, quality control procedures, and exclusion criteria. Without such reporting, the generalizability of neuroimaging findings remains limited, particularly in diverse populations.

Moving forward, it is essential for the field to develop and test targeted mitigation strategies to address these phenotypic impedances. This includes engineering improvements in optode design, better cap configurations for dense or curly hair, and novel preprocessing techniques that explicitly model or correct for signal loss due to pigmentation. Crucially, these solutions must be validated across a broad spectrum of hair and skin types to ensure equity in data acquisition and interpretation. By treating phenotypic signal barriers not as artifacts but as

solvable design and analytic challenges, the field of fNIRS can move toward more inclusive and scientifically rigorous practices.

6. Conclusion

This study demonstrates that task-based fNIRS reliably engages frontoparietal networks, producing effect sizes comparable to prior fMRI studies of reasoning. ROI-level analyses enhanced sensitivity to broad activation, while channel-level analyses captured finer spatial detail, particularly for condition-specific contrasts. Approximately nine blocks of data optimized the balance between signal stability and participant burden, though complex contrasts may require additional data. Functional connectivity emerged as a highly reliable metric across sessions and spatial scales, outperforming traditional activation measures and underscoring its utility for longitudinal and individualized research.

Group-level analyses revealed greater between-session stability for broad ROI-level contrasts. However, reliability at the individual level was more limited, with only select prefrontal channels and connectivity metrics showing potential. These findings highlight the value of combining ROI- and channel-level approaches and incorporating multiple reliability metrics, including amplitude-, pattern-, and variance-based measures, to capture the full temporal and spatial dynamics of fNIRS signals.

Importantly, signal quality was lower in parietal regions, contributing to reduced reliability and emphasizing the methodological challenge of maintaining consistent coverage across the scalp. The observed low reliability in parietal channels likely reflects these signal quality issues, which should be considered when interpreting results. Additionally, while the task was effective in eliciting reasoning-related activation, it may have lacked the contrast strength necessary to produce highly stable condition-specific effects. Future work should extend this framework to pediatric populations, where signal quality and participant variability pose greater challenges.

Finally, we emphasize the impact of hair and skin phenotypes on fNIRS signal quality. Darker pigmentation and hair were associated with greater signal loss, reinforcing the need for standardized phenotypic reporting and methodological innovations to improve data equity. Together, these findings support fNIRS as a promising tool for cognitive neuroscience, while identifying critical areas for refinement to enhance its reliability, inclusivity, and generalizability across diverse populations and study designs.

Limitations and Future Directions

Several limitations should be acknowledged when interpreting the current findings. As anticipated, signal quality was lower in parietal regions, which likely contributed to reduced reliability in these areas and underscores persistent challenges in achieving consistent coverage across the scalp. Additionally, the task manipulation employed, while well-validated for eliciting relational reasoning, may have been too subtle to generate highly reliable activation patterns, particularly for complex contrasts such as 2nd vs. 1st-order conditions. These results

are based on a single experimental paradigm and sample of healthy young adults; future work should extend this reliability framework to pediatric populations, where greater inter-subject variability and lower signal quality may be expected. To ensure reproducible results in such contexts, it is especially important to focus on the most robust analytic metrics, such as ROI-level functional connectivity. A critical avenue for future research involves addressing the impact of phenotypic variation on signal quality, including the well-documented impacts of darker or more textured hair and higher melanin content in skin. Although recent advances in hardware design have aimed to mitigate these disparities, systematic evaluation of their effectiveness is still urgently needed to ensure fNIRS can serve as an equitable and reliable tool in diverse populations.

Acknowledgements

This project was funded by a Jacobs Foundation Creating Impact Science Program (CRISP) fellowship (S.A.B.), an equipment award from the Berkeley Faculty Research Fund for the Biological Sciences (Professor Fei Xu and S.A.B., and University of California Berkeley internal research funds (K.J.). Research assistants Nick Zhang, Angela Ling, Nataly Lopez, Qi Guo, Kate Dellet, and Allison Chen assisted with data collection, and Julia Dil assisted with both data collection and a meta-analysis.

References

- Abdi, H. (2010). Holm's sequential bonferroni procedure. *Encyclop. Res. Des.*, 1, 1–8.
- Amaireh, A., Troller-Renfree, S., Chadwick, M., & Nelson, C. A. (2024). Caregiver executive functions are associated with infant visual working memory performance. *Infant and Child Development*, 33(2), e2432. <https://doi.org/10.1002/icd.2432>
- Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., & Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *NeuroImage*, 59(1), 36–47.
- Baldo, J. V., Bunge, S. A., Wilson, S. M., & Dronkers, N. F. (2010). Is relational reasoning dependent on language? A voxel-based lesion symptom mapping study. *Brain and Language*, 113(2), 59–64. <https://doi.org/10.1016/j.bandl.2010.01.004>
- Barker, J. W., Aarabi, A., & Huppert, T. J. (2013). Autoregressive model based algorithm for correcting motion and serially correlated errors in fNIRS. *Biomed. Opt. Express*, 4(8), 1366–1379.

Bazargani, N., Hillebrandt, H., Christoff, K., & Dumontheil, I. (2014). Developmental changes in effective connectivity associated with relational reasoning. *Human Brain Mapping*, 35, 3262–3276. <https://doi.org/10.1002/hbm.22394>

Beauchamp, M. S., Beurlot, M. R., Fava, E., Nath, A. R., Parikh, N. A., Saad, Z. S., Bortfeld, H., & Oghalai, J. S. (2011). The developmental trajectory of brain-scalp distance from birth through childhood: implications for functional neuroimaging. *PLOS ONE*, 6(9), e24981. <https://doi.org/10.1371/journal.pone.0024981>

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 57(1), 289–300.

Bennett, C., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Ann. N.Y. Acad. Sci.*, ISSN 0077-8923.

Bradford, D. E., DeFalco, A., Perkins, E. R., Carbajal, I., Kwasa, J., Goodman, F. R., ... & Joyner, K. J. (2024). Whose signals are being amplified? Toward a more equitable clinical psychophysiology. *Clinical Psychological Science*, 12(2), 237-252.

Brown, W. (1910). Some experimental results in the correlation of mental abilities 1. *British Journal of Psychology*, 1904-1920, 3(3), 296–322.

Bunge, S. A., Helskog, E. H., & Wendelken, C. (2009). Left, but not right, rostrolateral prefrontal cortex meets a stringent test of the relational integration hypothesis. *NeuroImage*, 46(1), 338–342. <https://doi.org/10.1016/j.neuroimage.2009.02.027>

Bunge, S. A., Wendelken, C., Badre, D., & Wagner, A. D. (2005). Analogical reasoning and prefrontal cortex: Evidence for separable retrieval and integration mechanisms. *Cerebral Cortex*, 15(3), 239–249. <https://doi.org/10.1093/cercor/bhh126>

Burgess, P., Crum, J., Pinti, P., Aichelburg, C., Oliver, D., Lind, F., Power, S., Swingle, E., Hakim, U., Merla, A., Gilbert, S., Tachtsidis, I., Hamilton, A., (2022). Prefrontal cortical activation associated with prospective memory while walking around a real-world street environment, *NeuroImage*, Volume 258, 2022, 119392, ISSN 1053-8119, <https://doi.org/10.1016/j.neuroimage.2022.119392>

Chenot, Q., Lepron, E., De Boissezon, X., & Scannella, S. (2021). Functional Connectivity Within the Fronto-Parietal Network Predicts Complex Task Performance: A fNIRS Study. *Front Neuroergon*, 2, 718176. doi: 10.3389/fnrgo.2021.718176.

Cho, S., Moody, T. D., Fernandino, L., Mumford, J. A., Poldrack, R. A., Cannon, T. D., Knowlton, B. J., & Holyoak, K. J. (2010). Common and dissociable prefrontal loci associated with component mechanisms of analogical reasoning. *Cerebral Cortex*, 20(3), 524–533. <https://doi.org/10.1093/cercor/bhp121>

Christoff, K., & Gabrieli, J. D. E. (2002). The frontopolar cortex and human cognition: Evidence for a rostrocaudal hierarchical organization within the human prefrontal cortex. *Psychobiology*, 28(2), 168–186. <https://doi.org/10.3758/BF03331976>

Christoff, K., Geddes, L., & Gabrieli, J. (2004). Evaluating Self-Generated Information: Anterior Prefrontal Contributions to Human Cognition. *Behavioral Neuroscience*, DOI: 10.1037/0735-7044.117.6.1161.

Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., & Gabrieli, J. D. (2001). Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *NeuroImage*, 14(5), 1136–1149. <https://doi.org/10.1006/nimg.2001.0922>

Christoff, K., Ream, J. M., Geddes, L. P. T., & Gabrieli, J. D. E. (2003). Evaluating self-generated information: Anterior prefrontal contributions to human cognition. *Behavioral Neuroscience*, 117(6), 1161–1168. <https://doi.org/10.1037/0735-7044.117.6.1161>

Cocchi, L., Halford, G. S., Zalesky, A., Harding, I. H., Ramm, B. J., Cutmore, T., Shum, D. H., & Mattingly, J. B. (2013). Complexity in relational processing predicts changes in functional brain network dynamics. *Cerebral Cortex*, 24(9), 2283–2296. <https://doi.org/10.1093/cercor/bhs222>

Crone, E. A., Wendelken, C., van Leijenhorst, L., Honomichl, R. D., Christoff, K., & Bunge, S. A. (2009). Neurocognitive development of relational reasoning. *Developmental Science*, 12(1), 55–66. <https://doi.org/10.1111/j.1467-7687.2008.00743.x>

Simone Cutini, S., Scatturin, P., Menon, E., Bisiacchi, P. S., Gamberini, L., Zorzi, M., Dell'Acqua, R., (2008) Selective activation of the superior frontal gyrus in task-switching: An event-related fNIRS study, *NeuroImage*, Volume 42, Issue 2, 2008, Pages 945-955, ISSN 1053-8119, <https://doi.org/10.1016/j.neuroimage.2008.05.013>.

de Rond, V., Gilat, M., D'Cruz, N., Hulzinga, F., Orban de Xivry, J. J., & Nieuwboer, A. (2023). Test-retest reliability of functional near-infrared spectroscopy during a finger-tapping and postural task in healthy older adults. *Neurophotonics*, 10(2), 025010. <https://doi.org/10.1117/1.NPh.10.2.025010>

DeYoung, C. G., Hilger, K., Hanson, J. L., Abend, R., Allen, T. A., Beaty, R. E., ... & Wacker, J. (2025). Beyond increasing sample sizes: Optimizing effect sizes in neuroimaging research on individual differences. *Journal of Cognitive Neuroscience*, 37(6), 1023-1034.

Dumontheil, I., Houlton, R., Christoff, K., & Blakemore, S. J. (2010). Development of relational reasoning during adolescence. *Developmental Science*, 13(6), F15–F24. <https://doi.org/10.1111/j.1467-7687.2010.01014.x>

Ebisch, S. J. H., Mantini, D., Romanelli, R., Tommasi, M., Perrucci, M. G., Romani, G. L., Colom, R., & Saggino, A. (2013). Long-range functional interactions of anterior insula and medial frontal cortex are differently modulated by visuospatial and inductive reasoning tasks. *NeuroImage*, 78, 426–438. <https://doi.org/10.1016/j.neuroimage.2013.03.008>

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

Eslinger, P. J., Blair, C., Wang, J., Lipovsky, B., Realmuto, J., Baker, D., Thorne, S., Gamson, D., Zimmerman, E., Rohrer, L., et al. (2009). Developmental shifts in fMRI activations during visuospatial relational reasoning. *Brain and Cognition*, 69(1), 1–10. <https://doi.org/10.1016/j.bandc.2008.04.010>

Ferrari, M., & Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *NeuroImage*, 63, 921–935.

Ferrer, E., Whitaker, K., & Bunge, S. (2014). Fronto-Parietal Network Reconfiguration Supports the Development of Reasoning Ability. *Cerebral Cortex*, 26(6), 2824–2836. <https://doi.org/10.1093/cercor/bhv050>.

Fiske, A., de Klerk, C., Lui, K. Y. K., Collins-Jones, L., Hendry, A., Greenhalgh, I., Hall, A., Scerif, G., Dvergsdal, H., & Holmboe, K. (2022). The neural correlates of inhibitory control in 10-month-old infants: A functional near-infrared spectroscopy study. *NeuroImage*, 257, 119241. <https://doi.org/10.1016/j.neuroimage.2022.119241>

Green, A. E., Fugelsang, J. A., Kraemer, D. J., Shamosh, N. A., & Dunbar, K. N. (2006). Frontopolar cortex mediates abstract integration in analogy. *Brain Research*, 1096(1), 125–137. <https://doi.org/10.1016/j.brainres.2006.04.024>

Green, A.E., Spiegel, K., Giangrande E., Weinberger, A., Gallagher, N., Turkeltaub, P., Thinking Cap Plus Thinking Zap: tDCS of Frontopolar Cortex Improves Creative Analogical Reasoning and Facilitates Conscious Augmentation of State Creativity in Verb Generation, *Cerebral Cortex*, Volume 27, Issue 4, April 2017, Pages 2628–2639, <https://doi.org/10.1093/cercor/bhw080>

Gu, Y., Miao, S., Han, J., Zeng, K., Ouyang, G., Yang, J., & Li, X. (2017). Complexity analysis of fNIRS signals in ADHD children during working memory task. *Scientific Reports*, 7, 829. <https://doi.org/10.1038/s41598-017-00965-4>

Halford, G. S., Wilson, W. H., & Philips, S. (1998). Processing Capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, 21, 803-865.

Hancock, A. S., Warren, C. M., Barrett, T. S., Bolton, D. A. E., & Gillam, R. B. (2023). Functional near-infrared spectroscopy measures of neural activity in children with and without developmental language disorder during a working memory task. *Brain and Behavior*, 13, e2895. <https://doi.org/10.1002/brb3.2895>

Haier, R. J., Siegel, B. V., Jr, MacLachlan, A., Soderling, E., Lottenberg, S., & Buchsbaum, M. S. (1992). Regional glucose metabolic changes after learning a complex visuospatial/motor task: a positron emission tomographic study. *Brain Research*, 570(1-2), 134-143.

Heilicher, M., Crombie, K. M., & Cisler, J. M. (2022). Test-Retest Reliability of fMRI During an Emotion Processing Task: Investigating the Impact of Analytical Approaches on ICC Values. *Frontiers in Neuroimaging*, 1. <https://doi.org/10.3389/fnimg.2022.859792>

Hempel, A., Giesel, F. L., Garcia, N. M., Armbruster, V. G., Wolff, M., Magnotta, V. A., Schröder, J. (2004). Plasticity of cortical activation after working memory training: a functional magnetic resonance imaging study. *NeuroImage*, 23(2), 748-756.

Hernandez, M. E., Holtzer, R., Chapina, J., Izzetoglu, M., Sliwinski, M., & Verghese, J. (2019). Distinct fNIRS-Derived HbO₂ Trajectories During the Course and Over Repeated Walking Trials. *Journals of Gerontology Series A*, 74(7), 1137-1145.

Hocke, L. M., Oni, I. K., Duszynski, C. C., Corrigan, A. V., Frederick, B. D., & Dunn, J. F. (2018). Automated processing of fNIRS data—A visual guide to the pitfalls and consequences. *Algorithms*, 11, 10.3390/a11050067.

Holmes, M., Aalto, D., & Cummine, J. (2024). Opening the dialogue: A preliminary exploration of hair color, hair cleanliness, light, and motion effects on fNIRS signal quality. *PLOS ONE*, 19(5), e0304356. <https://doi.org/10.1371/journal.pone.0304356>

Holyoak, K. J., & Monti, M. M. (2021). Relational Integration in the Human Brain: A Review and Synthesis. *J Cogn Neurosci*, 33(3), 341-356. doi: 10.1162/jocn_a_01619.

Huppert, T. J. (2016). Commentary on the statistical properties of noise and its implication on general linear models in functional near-infrared spectroscopy. *Neurophotonics*, 3, 10401.

Huppert, T. J., Hoge, R. D., Diamond, S. G., Franceschini, M. A., & Boas, D. A. (2006). A temporal comparison of BOLD, ASL, and NIRS hemodynamic responses to motor stimuli in adult humans. *NeuroImage*, 29, 368–382.

Jung, R., & Haier, R. (2007). The Parieto-Frontal Integration Theory (P-FIT) of intelligence: Converging neuroimaging evidence. *BEHAVIORAL AND BRAIN SCIENCES*, 30, 135–187. DOI: 10.1017/S0140525X07001185.

Karbach, J., & Kray, J. (2009). How useful is executive control training? Age differences in near and far transfer of task-switching training. *Dev Sci*, 12(6), 978-90. doi: 10.1111/j.1467-7687.2009.00846.x.

Karbach, J., Könen, T., & Spengler, M. (2017). Who Benefits the Most? Individual Differences in the Transfer of Executive Control Training Across the Lifespan. *J Cogn Enhanc*, 1, 394–405. <https://doi.org/10.1007/s41465-017-0054-z>

Kelly, A. M., & Garavan, H. (2005). Human Functional Neuroimaging of Brain Changes Associated with Practice. *Cerebral Cortex*, 15(8), 1089-1102.

Klein, J. H. (2024). Optimizing spatial specificity and signal quality in fNIRS-based real-time applications. *Frontiers in Neuroergonomics*, 5, 1286586.
<https://doi.org/10.3389/fnrgo.2024.1286586>

Klem, G. H., Lüders, H. O., Jasper, H. H., & Elger, C. (1999). The ten-twenty electrode system of the international federation. The international federation of clinical neurophysiology. *Electroencephalography and Clinical Neurophysiology. Supplement*, 52, 3–6.

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

Lloyd-Fox, S., Blasi, A., & Elwell, C. E. (2010). Illuminating the developing brain: The past, present and future of functional near infrared spectroscopy. *Neuroscience and Biobehavioral Reviews*, 34, 269–284.

Mandrick, K., Peysakhovich, V., Rémy, F., Lepron, E., & Causse, M. (2017). Neural Efficiency Index as a New Assessment of Mental Workload: A fNIRS Study. *Scientific Reports*, 7, 5222.

Meidenbauer, K. L., Choe, K. W., Cardenas-Iniguez, C., Huppert, T. J., & Berman, M. G. (2021). Load-dependent relationships between frontal fNIRS activity and performance: A data-driven PLS approach. *NeuroImage*, 230, 117795, ISSN 1053-8119,
<https://doi.org/10.1016/j.neuroimage.2021.117795>.

Mollison, A. A., Rodriguez, E. R., Metzger, A., & Shader, M. J. (2026). *Group-level test–retest reliability assessment using systemic physiology–augmented functional near-infrared spectroscopy during a passive-listening task*. *Neurophotonics*, 13(1), 015005

Montero-Hernandez, S., & Pollonini, L. (2019). QT-NIRS (Quality Testing of Near Infrared Scans) [Software]. University of Houston. <https://github.com/lpollonini/qt-nirs>

Morais, G. A. Z., Balardin, J. B., & Sato, J. R. (2018). fNIRS Optodes' Location Decider (fOLD): a toolbox for probe arrangement guided by brain regions-of-interest. *Scientific Reports*, 8, 3341.

Moriguchi, Y., & Hiraki, K. (2013). Prefrontal cortex and executive function in young children: A review of NIRS studies. *Frontiers in Human Neuroscience*, 7, Article 867.
<https://doi.org/10.3389/fnhum.2013.00867>

Nittel, C., Hohmann, D. M., Jansen, A., Sommer, J., Krauß, R., Völk, M., Kamp-Becker, I., Weber, S., Becker, B., Stroth, S. (2025). Test–retest reliability of functional near infrared spectroscopy during tasks of inhibitory control and working memory. *Psychiatry Research: Neuroimaging*, 350, 111993, ISSN 0925-4927.
<https://doi.org/10.1016/j.psychresns.2025.111993>.

Noble, S., Scheinost, D., & Constable, R. T. (2021). A guide to the measurement and interpretation of fMRI test-retest reliability. *Current Opinion in Behavioral Sciences*, 40, 27-32, ISSN 2352-1546. <https://doi.org/10.1016/j.cobeha.2020.12.012>.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51, 195–203.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*. 10.3758/s13428-018-01193-y.

Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., & Burgess, P. W. (2018). The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Annals of the New York Academy of Sciences*.

Pinti, P., Scholkmann, F., Hamilton, A., Burgess, P., & Tachtsidis, I. (2019). Current status and issues regarding pre-processing of fNIRS neuroimaging data: an investigation of diverse signal filtering methods within a general linear model framework. *Front. Hum. Neurosci.*, 12, 505. 10.3389/fnhum.2018.00505.

Plichta, M. M., Heinzel, S., Ehli, A.-C., Pauli, P., & Fallgatter, A. J. (2006). *Event-related functional near-infrared spectroscopy (fNIRS): Are the measurements reliable?* *NeuroImage*, 31(1), 116–124. <https://doi.org/10.1016/j.neuroimage.2005.12.008>

Santosa, H., Fishburn, F., Zhai, X., & Huppert, T. J. (2019). Investigation of the sensitivity-specificity of canonical- and deconvolution-based linear models in evoked functional near-infrared spectroscopy. *Neurophotonics*, 6(2), 025009.

Santosa, H., Zhai, X., Fishburn, F., & Huppert, T. (2018). The NIRS brain AnalyzIR toolbox. *Algorithms*, 11, 73.

Schwarze, S. A., Laube, C., Khosravani, N., Lindenberger, U., Bunge, S. A., & Fandakova, Y. (2025). Intensive task-switching training and single-task training differentially affect behavioral and neural manifestations of cognitive control in children. *Cereb Cortex*, 35(5), bhaf103. doi: 10.1093/cercor/bhaf103.

Shokri-Kojori, E., Motes, M. A., Rypma, B., & Krawczyk, D. C. (2012). The network architecture of cortical processing in visuo-spatial reasoning. *Scientific Reports*, 2, 411. <https://doi.org/10.1038/srep00411>

Simmons, A. S., Abdurkmonova, G., Taylor-Robinette, E. K., & Romeo, R. R. (2025). *Developing best practices for inclusion in fNIRS research: Equity for participants with Afro-textured hair* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/scbrq>

Sinfield, V. C., Aaker, D., Metzger, A., & Shader, T. M. (2025). Intra-subject test–retest reliability for auditory-evoked functional near-infrared spectroscopy responses: Effects of systemic physiology correction. *Neurophotonics*, 12(1), 015015. <https://doi.org/10.1117/1.NPh.12.1.015015>

Skau, S., Helenius, O., Sundberg, K., Bunketorp-Käll, L., & Kuhn, H.-G. (2022). Proactive cognitive control, mathematical cognition and functional activity in the frontal and parietal cortex in primary school children: An fNIRS study. *Trends in Neuroscience and Education*, 28, 100180. <https://doi.org/10.1016/j.tine.2022.100180>

Smith, E., Anderson, A., Thurm, A., Shaw, P., Maeda, M., Chowdhry, F., Chernomordik, V., & Gandjbakhche, A. (2017). Prefrontal activation during executive tasks emerges over early childhood: Evidence from functional Near Infrared Spectroscopy. *Developmental Neuropsychology*, 42(4), 253–264. <https://doi.org/10.1080/87565641.2017.1318391>

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3(3), 271.

Starr, A., Leib, E., Younger, J., Uncapher, M., & Bunge, S. (2022). Relational thinking: An overlooked component of executive functioning. *Developmental Science*, 26. [10.1111/desc.13320](https://doi.org/10.1111/desc.13320).

Strangman, G., Franceschini, M. A., & Boas, D. A. (2003). Factors affecting the accuracy of near-infrared spectroscopy concentration calculations for focal changes in oxygenation parameters. *NeuroImage*, 18, 865–879.

Tachtsidis, I., & Scholkmann, F. (2016). False positives and false negatives in functional near-infrared spectroscopy: Issues, challenges, and the way forward. *Neurophotonics*, 3(3), 031405. <https://doi.org/10.1117/1.NPh.3.3.031405>

Tong, Y., Lindsey, K. P., & deB Frederick, B. (2011). Partitioning of physiological noise signals in the brain with concurrent near-infrared spectroscopy and fMRI. *Journal of Cerebral Blood Flow & Metabolism*, 31, 2352–2362.

Vendetti, M. S., & Bunge, S. A. (2014). Evolutionary and developmental changes in the lateral frontoparietal network: A little goes a long way for high-level cognition. *Neuron*, 84(5), 906–917. <https://doi.org/10.1016/j.neuron.2014.10.038>

Volle, E., Gonen-Yaacovi, G., Costello, A. D. L., Gilbert, S. J., & Burgess, P. W. (2011). The role of rostral prefrontal cortex in prospective memory: A voxel-based lesion study. *Neuropsychologia*, 49(8), 2185–2198. <https://doi.org/10.1016/j.neuropsychologia.2011.05.015>

Waechter, R. L., Goel, V., Raymont, V., Kruger, F., & Grafman, J. (2013). Transitive inference reasoning is impaired by focal lesions in parietal cortex rather than rostrolateral prefrontal cortex. *Neuropsychologia*, 51(3), 464–471. <https://doi.org/10.1016/j.neuropsychologia.2012.11.012>

Watson, C. E., & Chatterjee, A. (2012). A bilateral frontoparietal network underlies visuospatial analogical reasoning. *NeuroImage*, 59(3), 2831–2838. <https://doi.org/10.1016/j.neuroimage.2011.09.030>

Weibley, H., Di Filippo, M., Liu, X., Lazenby, L., Goscha, J., Ferreira, A., Muscalu, L., & Rader, N. (2021). fNIRS monitoring of infant prefrontal cortex during crawling and an executive functioning task. *Frontiers in Behavioral Neuroscience*, 15, 675366. <https://doi.org/10.3389/fnbeh.2021.675366>

Wendelken, C., & Bunge, S. A. (2010). Transitive inference: Distinct contributions of rostralateral prefrontal cortex and the hippocampus. *Journal of Cognitive Neuroscience*, 22(5), 837–847. <https://doi.org/10.1162/jocn.2009.21232>

Wendelken, C., Chung, D., & Bunge, S. (2011). Rostrolateral Prefrontal Cortex: Domain-General or Domain-Sensitive? *Human Brain Mapping*, 33(8), 1952–1963.

Wendelken, C., Chung, D., & Bunge, S. A. (2012). Rostrolateral prefrontal cortex: Domain-general or domain-sensitive? *Human Brain Mapping*, 33(8), 1952–1963. <https://doi.org/10.1002/hbm.21336>

Wendelken, C., Nakhachenko, D., Donohue, S. E., Carter, C. S., & Bunge, S. A. (2008). “Brain is to thought as stomach is to ??”: Investigating the role of rostralateral prefrontal cortex in relational reasoning. *Journal of Cognitive Neuroscience*, 20(4), 682–693. <https://doi.org/10.1162/jocn.2008.20055>

Wendelken, C., O’Hare, E. D., Whitaker, K. J., Ferrer, E., & Bunge, S. A. (2011). Increased functional selectivity over development in rostralateral prefrontal cortex. *Journal of Neuroscience*, 31(47), 17260–17268. <https://doi.org/10.1523/JNEUROSCI.1193-10.2011>

Wright, S. B., Matlen, B. J., Baym, C. L., Ferrer, E., & Bunge, S. A. (2008). Neural correlates of fluid reasoning in children and adults. *Frontiers in Human Neuroscience*, 1, 8. <https://doi.org/10.3389/neuro.09.008.2007>

Xu, G., Huo, C., Yin, J., Zhong, Y., Sun, G., Fan, Y., Wang, D., & Li, Z. (2023). Test–retest reliability of fNIRS in resting-state cortical activity and brain network assessment in stroke patients. *Biomedical Optics Express*, 14(8), 4217–4236. <https://doi.org/10.1364/BOE.487399>

Yeung, M. K., & Han, Y. M. Y. (2023). Changes in task performance and frontal cortex activation within and over sessions during the n-back task. *Sci Rep*, 13, 3363. <https://doi.org/10.1038/s41598-023-30552-9>

Yücel, M. A., Anderson, J. E., Rogers, D. J., Hajirahimi, P., Farzam, P., Gao, Y., Kaplan, R. I., Braun, E. J., Mukadam, N., Duwadi, S., Carlton, L., Beeler, D., Butler, L. K., Carpenter, E., Girnis, J., Wilson, J., Tripathi, V., Zhang, Y., Sorger, B., von Lühmann, A., Somers, D. C., Cronin-Golomb, A., Kiran, S., Ellis, T. D., & Boas, D. A. (2024). Inclusivity in fNIRS studies:

Quantifying the impact of hair and skin characteristics on signal quality with practical recommendations for improvement. *bioRxiv*. <https://doi.org/10.1101/2024.10.28.620644>

Yücel, M. A., Lühmann, A. V., Scholkmann, F., Gervain, J., Dan, I., Ayaz, H., Boas, D., Cooper, R. J., Culver, J., Elwell, C. E., Eggebrecht, A., Franceschini, M. A., Grova, C., Homae, F., Lesage, F., Obrig, H., Tachtsidis, I., Tak, S., Tong, Y., Torricelli, A., Wabnitz, H., & Wolf, M. (2021). Best practices for fNIRS publications. *Neurophotonics*, 8(1), 012101. <https://doi.org/10.1117/1.NPh.8.1.012101>

Zhilin Shu, Jingchao Wu, Haitao Li, Jinrui Liu, Jiewei Lu, Jianeng Lin, Siquan Liang, Jialing Wu, Jianda Han, & Ningbo Yu. (2023). fNIRS-based functional connectivity signifies recovery in patients with disorders of consciousness after DBS treatment. *Clinical Neurophysiology*, 147, 60-68. ISSN 1388-2457, <https://doi.org/10.1016/j.clinph.2022.12.011>.

Zou, J., Yin, Y., Lin, Z., & Gong, Y. (2023). The analysis of brain functional connectivity of post-stroke cognitive impairment patients: an fNIRS study. *Front Neurosci*, 17, 1168773. doi: 10.3389/fnins.2023.1168773.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Data available on request. Please contact the corresponding author.

Highlights

- We measured within- and between-session reliability of fNIRS metrics
- Collecting more data yielded diminishing returns in effect size and precision
- There were tradeoffs to aggregating channel data into regions of interest
- General task activation was more reliable than a specific task contrast
- Functional connectivity showed extremely high test–retest reliability