Introduction to Cognitive Testing Data in the *All of Us* Research Program

Authors: Roger Strong and Amy Price

Table of Contents

Document Overview

This document serves as an overview of the cognitive testing data available through the *All of Us* Research Program. Our primary objective is to provide researchers with a thorough understanding of the cognitive task data structure, the categories of data produced, and key considerations for utilizing this data in various analytical contexts.

We aim to equip researchers with essential details, relevant links, and valuable resources to enhance your understanding of the cognitive task data within the *All of Us* Researcher Workbench. While we cannot offer specific instructions on data pre-processing and analysis—as these methods vary depending on study design and research objectives—we provide an overview of the available data elements to help initiate research using the *All of Us* cognitive task data.

This guide covers:

1. The structure and methodology of cognitive testing in the All of Us Research Program
2. A summary of the various categories of cognitive data collected
3. Important considerations for applying this data in different analytical scenarios
4. An overview of the cognitive task data elements accessible through the Researcher Workbench

We intend for this document to serve as a valuable starting point and reference tool as you begin incorporating cognitive task data into your studies on the Researcher Workbench. After

you finish reading this introduction to cognitive testing data, you can access each task in the User Support Hub Article Overview of Exploring the Mind Data in the Researcher Workbench.

What is Cognitive Testing?

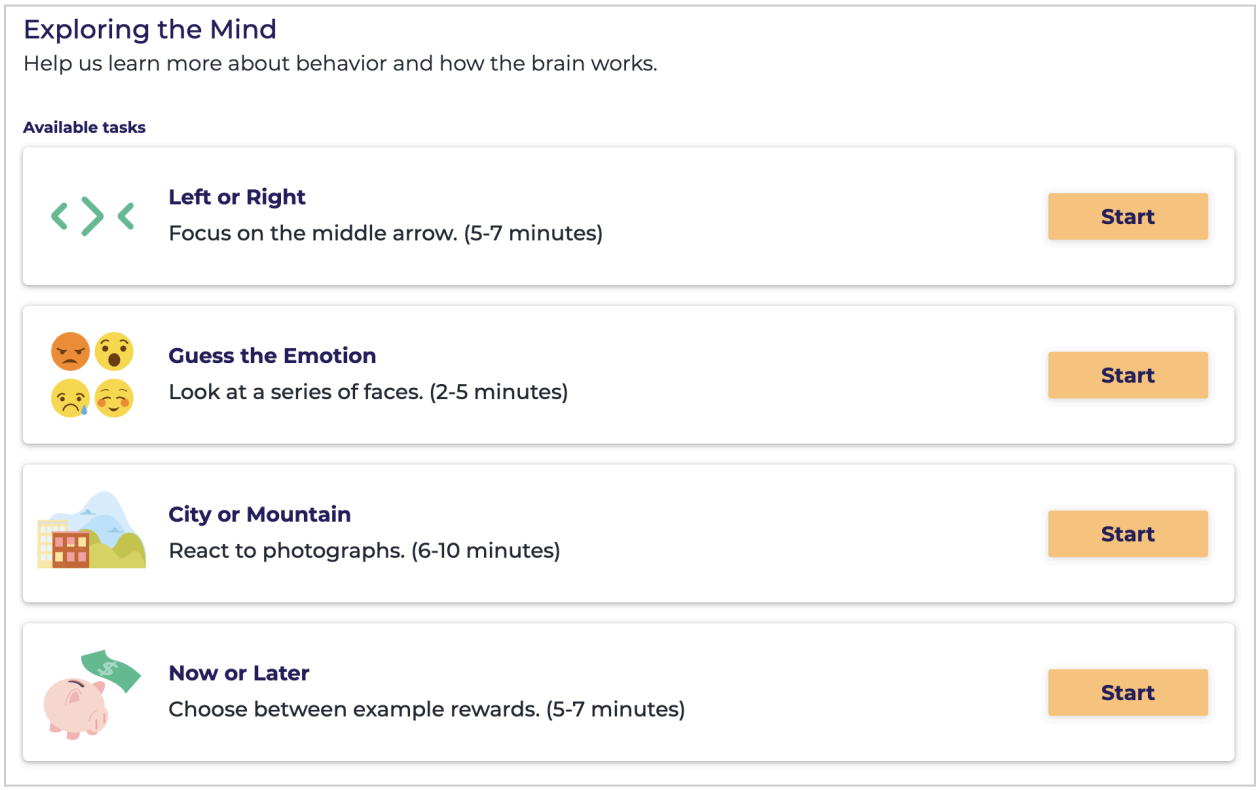
Human cognition comprises a complex combination of cognitive domains, including but not limited to domains like attention, memory, and executive functioning. *Cognitive tests* measure performance on one or more cognitive domains using an individual's responses to stimuli (e.g., pictures, words, sounds) while completing a task. The *All of Us* Research Program began providing four types of cognitive tests remotely to all eligible participants[1] in September 2023[2] that focus on measuring cognitive control, sustained attention, social facial recognition, and reward valuation. These tests were selected in collaboration with the National institute of Mental Health (NIMH) using the Research Domain Criteria (RDoC) framework (Cuthbert 2022; Cuthbert & Kozak, 2013; National Institute of Mental Health, 2023). The RDoC framework offers a set of principles for investigating mental disorders, unconstrained by diagnostic categories, which seeks to explain dimensions of functioning that span the full range of human behavior. RDoC promotes the use of multiple approaches, from genetic and molecular data to behavioral and self-reported data. The selected behavioral tasks tap into functional domains that can be disrupted in multiple disorders. These tests have previously been validated for remote, unsupervised administration on *TestMyBrain.org,* a web-based cognitive testing platform that allows large-scale collection of data, with similar data quality to traditional in-person cognitive testing (Germine et al., 2012). This software is supported by the 501c3 nonprofit the Many Brains Project (manybrains.net) and McLean Hospital.

The set of four game-like cognitive tests are referred to as the 'Exploring the Mind'[3] tasks to participants.[4] Participants are eligible to complete the Exploring the Mind tasks once they have completed the program's first three baseline surveys.[5] Once those surveys are complete, participants can log into the Exploring the Mind task gallery from the *All of Us* participant portal to complete any of the four tasks available in any order they choose (see Figure 1).[6] They are free to complete the tasks on a device type of their choosing (i.e., desktop, laptop, smartphone, tablet). Basic metadata from the device type and response method used (e.g., mouse click, touchscreen) are recorded and available for researchers (see below).

---

[1] Participants were eligible once they had completed the Basics survey, the Lifestyle survey, and the Overall Health survey.
[2] The tasks were first piloted with a small subset of *All of Us* participants from December 2022 to September 2023.
[3] This name was selected by the *All of Us* participant ambassadors.
[4] *All of Us* September 2023 newsletter, NIMH news release
[5] The Basics survey, the Lifestyle survey, and the Overall Health survey
[6] The order of task appearance in the task gallery was counterbalanced across all participants. Once the participant completes a task, the participant must wait until 30 days have elapsed for that specific task to become available again in the Exploring the Mind task gallery. This delay was intended to minimize the influence of frequent repeated attempts on performance.

Figure 1: Exploring the Mind Task Gallery



Cognitive Testing Structure

At the start of a cognitive test, a *participant* (someone completing a cognitive test) views instructions explaining what they must do to complete the test. Following the instructions, the participant completes a series of *trials* (sometimes also referred to as *items* or *questions*). Each trial begins with a stimulus (e.g., a picture, word, or sound) being presented to the participant, and each trial ends when the participant makes a response to the stimulus.[7] Following each trial, the characteristics of the stimulus and the participant's response are recorded. Some cognitive tests have *practice trials*, which allow the participant to experience an example of the test structure, prior to beginning the actual test. The test ends once all the trials composing the test are completed.[8]

As an example, consider a cognitive test measuring emotion recognition. At the start of the test, a participant views instructions reading "Click the emotion that best matches the face." The

[7] Details of the trial design including max response time and allowable response options are provided in each of the individual task articles.
[8] Practice trial data is available in the trial-level data, but does not contribute to any of the full test summary outcomes.

participant then completes 48 test trials. On each trial, the participant views an image of an actor depicting one of four emotions: (1) anger, (2) fear, (3) sadness, or (4) happiness. Each trial ends when the participant answers which of the four emotions they believe is depicted, by clicking a button on the screen corresponding to that emotion. Following each trial, information about the stimulus and the participant's response is recorded, including the emotion actually depicted by the actor in the photo, the emotion the participant answered as a response, the accuracy of the participant's response (i.e., whether they selected the correct emotion), and the participant's reaction time (the time elapsed between the image appearing and the button press). After the participant completes all the trials, data summarizing the participant's overall test performance (e.g., overall accuracy and average reaction time) is computed and recorded. Additionally, data contextualizing the participant's testing experience is recorded (e.g., time of test completion). See the following section for a summary of the different categories of data recorded for cognitive tests.

Cognitive Test Data

Categories of Cognitive Test Data available in the Researcher Workbench

The following sections describe the three main categories of cognitive test data available for each task: trial-level data, summary scores, and metadata. Each of these data categories may contain multiple data types (e.g., numeric values, text strings, boolean values).

The data dictionary linked here provides the full list of variables and their definitions for each task and each category.

Trial-level data
Trial-level data describes the characteristics of both (1) the stimulus the participant observed on each trial and (2) the participant's response on each trial.

Summary scores
At the end of a cognitive test, data obtained from the participant's responses to each individual trial (trial-level data) are aggregated to produce *summary scores* (sometimes also referred to as *outcomes*). In contrast to individual trial data, each summary score reflects performance across a combination of trials, often the entire test. Example summary scores might include (1) the proportion of trials where the participant made the correct response and (2) the participant's average reaction time across all trials. Summary scores can be computed from the trial-level data, but are nevertheless provided separately to allow researchers immediate access to scores that reflect overall performance on the test. Performance on practice trials is **never** included in the calculation of summary scores.

Metadata

Metadata refers to data that contextualizes the data generated from a participant's responses during a cognitive test, but is not directly generated from those responses. Metadata might include information about when the test was completed (e.g., start time and language of test administration) and characteristics of the participant's testing device (e.g., screen width, operating system).

Consideration of How and When to Use Different Data Categories

Researchers most often use summary scores when analyzing cognitive test data, as these scores reflect overall performance across the entire testing session, and thus are a better index of cognitive performance than performance on any individual trial of the test. Summary scores can be used for many types of analyses, including:

- determining how well or poorly a participant performed relative to other participants who completed the test
- characterizing whether variation in test performance is associated with variation in another variable (e.g., does performance on test A correlate with performance on test B?)
- testing whether test performance varies for an individual at different testing timepoints (e.g., before and after taking medication)
- removing participants who experienced quality control violations on a particular test (see *Quality Control Flags* section below)

Although summary scores are usually most useful and commonly used for characterizing cognitive test performance, researchers may wish to access trial-level data in certain situations, including:

- computing additional summary scores that have not already been provided (e.g., capturing overall trial-to-trial variability)
- analyzing whether performance varies throughout the course of the test (e.g., is performance better in the second half of the test than in the first)
- performing cognitive modeling that requires individual trial data (e.g. drift-diffusion modeling)
- computing the reliability of performance variation across trials within the test (e.g., split-half reliability)
- performing item-based psychometric analyses (e.g., for building an item-response theory model)
- performing analyses that exclude individual trials based on quality control violations (e.g., computing summary scores after excluding trials with implausibly fast reaction times)
- assessing performance on practice trials

Researchers may consider using metadata in combination with summary scores or trial-level data during data analysis for a variety of reasons:

- classifying devices used to complete cognitive tests. Some classification categories researchers might consider include:
    - classifying devices as “large” versus “small”
        - Given the landscape of devices in 2023, the Many Brains Project recommends classifying devices into two size categories: “small” for devices with *screen_width* less than 1024 AND *screen_height* less than 768, and otherwise classifying devices as “large”. Researchers may consider using other cutoffs based on their expertise, the current device landscape, and the research question.
    - classifying devices by the response input used (e.g., mouse, keyboard, or touch)
    - classifying devices by a combination of size and input type
- analyzing whether performance variation occurs due to factors that aren’t directly related to cognition, such as:
    - device used during testing: is test performance worse for participants who use smaller devices (e.g., a cell phone) than for participants who use larger devices (e.g., laptops and desktop computers)?
    - time of data collection: is performance worse for participants who complete testing late at night than during the middle of the day?
- limiting analyses to only participants who are completing the test for the first time
- limiting analyses to a certain time range
- limiting analyses to data collected from a particular version of a test

Data Quality Assessments

Reliability

An important characteristic of a cognitive test is its reliability. Measures of reliability assess whether a test yields a consistent outcome for a participant. A task measure has high reliability if it produces the same scores, or the same ordering of scores, for participants in a single testing session or across multiple testing sessions. If a task measure has poor reliability, any performance differences between participants (or within a participant across time) will not be meaningful, and therefore researchers will be unable to find reliable associations between performance on the test and other measures (Brysbaert, 2024; Zorowitz & Niv, 2023). Therefore, computing the reliability of performance differences in a given sample is an important first step in any analysis of cognitive testing data.

There are many possible approaches to calculating a test’s reliability (Revelle & Condon, 2019). When testing occurs in a single session, a measure of *internal consistency* such as Cronbach’s alpha (Taber, 2018) or split-half reliability (Pronk et al., 2022) is typically computed. There is no definitive benchmark for what constitutes acceptable reliability (Taber, 2018), but reporting a test’s reliability allows better interpretation of research results. Suggestions for using trial-level data to compute reliability are provided in each cognitive test’s User Support Hub article. When a participant completes a task more than once across sessions, test-retest reliability can be

computed. When learning effects are not an issue, test-rest reliability measures how consistently a participant completes the same task at different times. Because the majority of participants completed each of the tests a single time for this data release, measures of the internal consistency of a single testing session (e.g., split-half reliability, Cronbach's alpha) were computed and provided in each cognitive test's User Support Hub article instead of test-retest reliability.

Quality Control Flags

The goal of cognitive testing is to obtain an index of cognitive performance. However, data may not reflect an individual's actual cognitive performance for a variety of reasons, including (1) technical problems, (2) disengagement from the test (e.g., not paying attention or guessing randomly), (3) failure to understand test instructions, or (4) physical disabilities that affected the participant's performance that may be unrelated to cognition (e.g., disabilities affecting hand mobility or vision). To help researchers determine whether quality control violations indicative of invalid test performance have occurred, quality control flags are included in both trial-level data and summary score data. The quality control flags are unique to each task's design, and they are described in detail in each of the individual cognitive task articles in the User Support Hub. The quality control criteria provided for each test are intended to capture extreme deviations from what is typically seen in participants performing the tasks in a valid manner. Researchers must use their own judgment when determining whether flagged participants (or trials) should be excluded from analyses. Researchers may want to factor in other participant data from All of Us to make these decisions (e.g., survey data, electronic health records). Researchers may also consider implementing their own quality control criteria separately from these recommendations.

Quality control flags may be included in both trial-level data and summary score data:

- individual trials in trial-level data
    - The criteria for a trial being flagged for a quality control violation will differ for each test, and some tests may not be assessed for quality control at the individual trial level. Even when a subset of trials are flagged for a quality control violation, performance may still be valid for other trials and the test as a whole.
    - Flagged trials are still used when computing summary scores; however, based on a research design and question, researchers could consider using the trial-level data to recompute summary scores without the flagged trials.
- summary score data
    - Summary score fields beginning with "flagged_" are quality control flags that apply to performance on the entire test
    - When fields beginning with "flagged_" have a value of 1, this indicates that the participant had a quality control violation suggesting invalid performance on the test as a whole
    - When analyzing data, researchers should consider excluding participants who have any summary score quality control flag variables equal to 1

References

Brysbaert, M. (2024). Designing and evaluating tasks to measure individual differences in experimental psychology: a tutorial. *Cognitive Research: Principles and Implications*, *9*(1), 11. https://doi.org/10.1186/s41235-024-00540-2

Cuthbert, B. N. (2022). Research domain criteria (RDoC): progress and potential. *Current Directions in Psychological Science*, *31*(2), 107-114. https://doi.org/10.1177/09637214211051363

Cuthbert, B. N., & Kozak, M. J. (2013). Constructing constructs for psychopathology: The NIMH research domain criteria. *Journal of Abnormal Psychology, 122*(3), 928–937. https://doi.org/10.1037/a0034028

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*, 847-857. https://doi.org/10.3758/s13423-012-0296-9

National Institute of Mental Health. (2024). Research Domain Criteria (RDoC). https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc (accessed Sep 10, 2024).

Pronk, T., Molenaar, D., Wiers, R. W., & Murre, J. (2022). Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. *Psychonomic Bulletin & Review, 29*(1), 44-54. https://doi.org/10.3758/s13423-021-01948-3

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395-1411. https://doi.org/10.1037/pas0000754

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*, 1273-1296. https://doi.org/10.1007/s11165-016-9602-2

Zorowitz, S. & Niv, Y. (2023). Improving the Reliability of Cognitive Task Measures: A Narrative Review. *Biol Psychiatry Cogn Neurosci Neuroimaging*, *8(8)*, 789-797. https://doi.org/10.1016/j.bpsc.2023.02.004