Spring 5-20-2022

Active Testing of Executive Functions: Toward More Efficient and Equitable Individual Behavioral Modeling

Mariluz Rojo Domingo

Recommended Citation

Rojo Domingo, Mariluz, "Active Testing of Executive Functions: Toward More Efficient and Equitable Individual Behavioral Modeling" (2022). *McKelvey School of Engineering Theses & Dissertations*. 703. https://openscholarship.wustl.edu/eng_etds/703

WASHINGTON UNIVERSITY IN ST. LOUIS

McKelvey School of Engineering
Department of Biomedical Engineering

Thesis Examination Committee:
Dennis Barbour, Chair
Jacob Gardner
Jason Hassenstab

Active Testing of Executive Functions: Toward More Efficient and Equitable Individual Behavioral Modeling

by
María De La Luz Rojo Domingo

A thesis presented to the McKelvey School of Engineering
of Washington University in St. Louis in partial fulfillment of the
requirements for the degree of
Master of Science

May 2022
St. Louis, Missouri

Table of Contents

List of Figures

Acknowledgments

First and foremost, my thesis would not have been possible without the help of my research advisor: Dr. Dennis Barbour. I would like to express my heartfelt appreciation for his contribution to my life and career. Thanks to his guidance and mentorship, I grew as a researcher, but his impact on me over the past two years was far greater than a purely academic one. He believed in me enough to invite me to join his lab and I was privileged to work under his supervision. Every step of the way, he provided me with the tools to face countless challenges, patiently helped me improve my skills, and unconditionally supported me through all the difficulties. Next, I would like to express my deep appreciation to my second research mentor, Dr. Jacob Gardner. His brilliant ideas and immense knowledge were invaluable in navigating unexplored scientific questions. His passion for pushing the limits of knowledge was an inspiration in my research endeavor, and I am sincerely grateful for all his insightful feedback. I would also like to thank the members of the Barbour lab, especially, Pranav Maddula, Quinn Wai Wong, Álvaro Grande, Daniel Fu, Michael Guo and Ethan Zheng for their contribution to this work. Last but not least, I would like to thank my parents for their love, sacrifices and endless support not only throughout the completion of this thesis, but throughout my life. Words cannot express how blessed I am to be your daughter.

María De La Luz Rojo Domingo

Washington University in St. Louis

May 2022

Dedicated to my parents, for their unwavering support.

ABSTRACT OF THE THESIS

Active Testing of Executive Functions: Toward More
Efficient and Equitable Individual Behavioral Modeling

by

María De La Luz Rojo Domingo

Master of Science in Biomedical Engineering

Washington University in St. Louis, 2022

Professor Dennis Barbour, Chair

Inferences about executive functions are commonly drawn through serial administration of various individual assessments that often take a long time to complete and cannot capture complex trends across multiple variables. In an attempt to improve upon current methods used to estimate latent brain constructs, this thesis makes two primary contributions to the field of behavioral modeling. First, it brings attention to sequential designs for more efficient diagnostic testing of fluctuations in executive functions with respect to a baseline level. It was shown that a sequential framework was successfully capable of detecting significant differences in cognitive performance more rapidly than conventional fixed approaches. Second, it introduces a scalable Gaussian Process estimator that can build individual psychometric models of task performance without requiring prohibitive amounts of data. This probabilistic machine learning classifier was capable of obtaining fully predictive models of working memory capacity person by person with high confidence.

Chapter 1: Introduction

Executive functions are at the center of cognitive psychology and comprehend the basic brain processes of working memory, cognitive flexibility and inhibitory control (Miyake et al., 2000). This core triad of executive functions is constituted by high-level cognitive processes that play a vital role in our everyday mental and physical activities (Diamond, 2013). Executive functions provide a foundation for higher-order processes such as reasoning, problem-solving and decision making (Viana-Sáenz et al., 2020). Considering that executive functions are the cornerstone of learning and academic success (Zelazo et al., 2016), rigorous short- and long-term characterization of improvements and fluctuations in these neurocognitive skills is essential. Neuroscience researchers typically employ quantitative measures of executive functions in order to gain insight into a subject's brain states. Performance-based numerical measures can be useful to examine both inter- and intra-individual changes in the manifestation of executive functions across time. Unfortunately, with current behavioral assessments, it should be recognized that a number of challenges persist to performing thorough studies of the relationships among executive functions, as well as their variability. To this day, multidimensional testing in perceptual or cognitive domains predominantly consists of serial administration of lengthy unidimensional tests that constitute a test battery. Refinement of traditional executive function testing procedures is indispensable to lay the groundwork for quantitatively predicting learning outcomes person by person more efficiently.

There exist numerous shortcomings in traditional cognitive and perceptual methods, such as the fact that current unidimensional tests within a battery broadly apply fixed designs with no active testing involved. Unfortunately, the lack of flexibility in one-size-fits-all assessments typically translates into excessive data requirements for accurate modeling, as there is no possibility of actively shortening test duration for any individual. In view of this major drawback of standardized behavioral tests, it is timely to formalize and extend the concept of variable termination inherent to adaptive procedures. By means of a sequential testing design, this work seeks to actively test whether changes in task performance have occurred from one session to another of the same task.

Considering that hypothesis testing generally entails low computational complexity, this study supports the notion that it is feasible to carry out multiple interim analyses before the end of a test session to more rapidly detect differences in executive functions and stop collecting data before reaching a maximum number of observations. In this thesis, it is demonstrated that relatively simple and straightforward frequentist tools can suffice to effectively tackle the problem of reducing total trial count in individual cognitive tests, while still controlling for error rates. It is also worth emphasizing that frequentist statistics, which are well-established tools that are commonly used in scientific inquiries, are the backbone of the sequential framework proposed. As a result, this method of conducting assessments does not introduce new techniques, but simply combines them in a way that exploits the benefits of statistical analyses. To promote the widespread use of sequential analyses in a variety of behavioral applications, this study sheds light on the great potential they offer to reduce test duration in the diagnosis of significant changes in cognitive performance.

Additionally, a major limitation in psychometrics is that rigorous estimation procedures for complex problems in higher dimensions are neither standardized nor well-established. The fact that unidimensional assessments fail to exploit and quantify the interactions between variables related to several executive functions cannot be ignored. Indeed, individual task performance measures involve a mixture of executive function processes, as it turns out to be difficult to isolate an index of a single functional process in any particular executive function task (Hughes & Graham, 2002). As a result, the fundamental failure of conventional approaches to capture nonlinear, multidimensional trends among psychological variables is substantially inefficient and makes scalability to higher dimensions unattainable.

Another source of concern is that current estimation methods combine measurements across individuals within a group to construct cognitive models. Although this method is useful for reducing noise and inferring similarities in a population, referencing individuals against each other and norming raw measurements introduces systematic errors and test bias (Reynolds & Suzuki, 2012). For instance, adaptive staircase designs are a widespread threshold-seeking procedure to summarize task performance and, when combined across a cohort, they allow researchers to infer group-level functions related to the underlying constructs of interest. The justification for this approach is that without referencing individuals against one another, the data requirements to draw individualized inference over complex latent constructs are excessive in higher dimensions due to the “curse of dimensionality” (Barbour, 2019; Feczko et al., 2019). The problem is that, especially in the case of heterogeneous cohorts of people, this population-based approach has a detrimental effect for minority or outlier examinees, whose performance tends to be misestimated with respect to the majority group (Reynolds & Suzuki, 2012). Thus, it is of paramount importance to design for improved estimation procedures that take a more equitable approach to estimate complex individual models. All in all, there is a dire need for innovative solutions for exploiting multidimensional feature spaces and retaining data efficiency for individual model training. To achieve this objective, this thesis puts forward a practical method for obtaining tractable complex individual models that can be trained with reasonable amounts of data applying modern estimation techniques.

To recapitulate, regarding the limitations that come with using traditional behavioral assessments of perceptual and cognitive processes, this thesis aims to improve upon current methods so as to promote more equitable and efficient assessments in cognitive psychology, neuroscience and education. Chapter 2 proposes a frequentist sequential testing statistical framework that can shorten test duration and lead to faster screening for changes in executive functions. Empirical results in non threshold-seeking cognitive tasks endorsed that even with hypothesis testing techniques only, it is possible to detect fluctuations between two test sessions with fewer observations than standard testing procedures. Next, chapter 3 establishes a sophisticated inference method that employs Bayesian machine learning tools to build a novel probabilistic estimator. This novel active learning Gaussian Process estimator, which is described in more detail below, was evaluated to quantify spatial working memory. This promising implementation, which is generalizable to other tasks, lays an essential foundation for incorporating additional executive functions beyond those from working memory capacity. Chapter 4 summarizes the conclusions of the two main contributions of this thesis to the field of cognitive neuroscience, and outlines future expansions of this line of research to continue to advance the field of behavioral modeling.

Chapter 2: Active Executive Function Difference Detection

Cognitive testing is essential for evaluating executive functions, but efforts are still needed to reduce test burden. Fortunately, by using more efficient data collection designs, shorter behavioral assessments can be developed. Scaling down data requirements in cognitive tests may also facilitate more frequent testing, allowing one to more easily determine if brain states have changed on any given day. Construct-valid tests designed to measure a given executive function can be informative to determine whether a fluctuation from baseline has occurred. Direct comparison of individual performance from one test session to another is one way to gauge intra-subject fluctuations in executive functions. A difference detection strategy can be applied to compare sets of responses from the same individual and the same assessment, then decide whether performance in one test session is significantly different from performance in another session. Based on task performance, the simplest way to make a quick diagnosis of significant changes in executive functioning is a binary output: "different" vs "same", or equivalently, "change" vs "no change". Hypothesis testing is a classical frequentist approach to make such determination by rejecting, or failing to reject, the null hypothesis: "inferring from performance in the cognitive test, there are no detectable fluctuations in executive function with respect to the baseline".

It is important to recognize that there are many different types of executive functioning assessments, so researchers should adapt the difference detection strategy to the nature of the task at hand, rather than searching for a one-size-fits-all solution. Overall, tasks can be largely classified as threshold-seeking or non threshold-seeking. Threshold-seeking assessments tend to have independent variables and correspond to models focused on specifying psychometric performance thresholds, such as the maximum number of items one can remember in a Corsi span task. On the other hand, non-threshold seeking approaches may not have independent variables and typically yield an output metric that summarizes performance on all responses, such as mean reaction time in a timing task. The non threshold-seeking cognitive assessments tackled in this chapter can be divided into two categories: timing models and accuracy models. The focus of this study is on the application of sequential analyses for difference detection in non threshold-seeking models that do not have a validated mechanism for adjusting task difficulty. Then, solutions for improving upon current psychometric function estimation procedures will be addressed in chapter 3.

2.1 Sequential Testing

Sequential analyses are far from being a new invention. As a matter of fact, statistical techniques to sequentially analyze data are well-established procedures that have a long history (Dodge & Romig, 1929). Testing items sequentially, one after another, is common practice in quality control in manufacturing and detection of anomalies in medical trials (Eggen, 1999; Spiegelhalter, 2003). To date, despite the popularity of sequential testing in fields such as medicine, one may find it surprising that this method is underutilized in most other research disciplines (Lakens et al., 2021). Due to the increased efficiency gains that a sequential design can offer (Neumann et al., 2017), it is worth exploring the potential for this approach to hypothesis testing in the field of cognition. Mainstream use of sequential testing as a screening tool in behavioral and psychological sciences could translate into detecting differences in executive functions in shorter periods of time. With this in mind, the goal in this chapter is to contrast sequential testing procedures with conventional non-sequential testing routines in the scope of detection of differences in executive functions relative to a cohort average or changes relative to an individual baseline.

In this study, task performance on a full test, where the maximum number of observations were presented, is used to establish a baseline executive function metric. Broadly, investing time to gather a large number of trials on the first testing session is justified by the fact that a reference model representing the underlying cognitive state as accurately as possible is desirable for a more efficient subsequent comparison to a new testing session. Departing from this baseline, the inference a researcher draws concerning the presence or absence of differences in brain states can be regarded as a dichotomy. With sequential tests, resolving this dichotomy in a more time-efficient manner is not only achievable, but also quite straightforward. The main idea of this technique is to establish a practical sampling plan to periodically run an interim analysis of the cumulative results, instead of waiting until the end to perform one final analysis after spending all the trials in the overall budget. An interim analysis can also be referred to as a "look", which means analyzing the data collected up to a point. In the context of detecting changes in executive functions, the purpose of the look is to determine whether there is enough evidence to endorse the presence of a difference in cognitive performance after a certain number of observations. Alternatively, it is possible to claim that the presence of an effect larger than the effect size of interest can be rejected because the observed effect at the interim inspection is much smaller than what is considered a large enough effect size for the particular cognitive test. In the latter case, the data collection process is interrupted for "futility", which means that early data shows insufficient promise of the presence of a difference in executive function. That is, it is either impossible or very unlikely for the final analysis to yield a significant p value.

Presumably, with every additional data point that a behavioral test collects, the amount of evidence in support of one of the two opposing hypotheses increases. The alternate and null hypotheses were respectively defined in this study as "there was a significant change in cognitive function" (H1), and "there was no significant change in cognitive function" (H0). Thus, after serial addition of each new sample, a hypothesis test can be performed to see if either conclusion can be reached with confidence. The testing session will be stopped if the null hypothesis can be rejected (e.g., a NHST), or because the alternate hypothesis can be rejected (e.g., an equivalence test). In practice, instead of inspecting the data after every single observation, performing an inspection after a small batch of samples reduces computational time. This sequential modality, known as group sequential design, was implemented in the non-threshold-seeking tests in this study. Concretely, the hypothesis tests that were performed after each batch of trials were Mann-Whitney U tests when the data streams contained reaction time information and binomial tests when the responses stored accuracy data.

It is intuitive that with more data collected, there is less uncertainty about the effect size and, thus, more information is available to researchers to draw inference from. In effect, the key issue in sequential designs lies in deciding when to terminate the investigation and validate a particular decision, contemplating the possibility that the number of trials required to draw a conclusion could certainly be less than the pre-established maximum number of trials planned. In general, fewer samples are needed to draw conclusions about large effects than small effects and forcing every participant to experience the same number of samples is inefficient when the effect sizes are likely to vary. At each interim analysis, based on the test responses gathered, one of the following three decisions is made: accept the null hypothesis, reject the null hypothesis, or continue testing. In the first case, the test is interrupted because enough evidence has been accumulated to determine that a significant change in executive function with respect to the previous session is extremely unlikely. In the second case, early stopping takes place at an interim look because there is sufficient evidence to resolve a change in executive function. In the third and last case, more samples are drawn and the test continues, because the information available after that particular number of observations is not quite enough to make a confident decision regarding whether a change in executive function has occurred. Therefore, test duration in sequential frameworks depends on performance at the individual level, unlike traditional testing procedures, where the number of observations collected is identical for all the subjects. For many individuals, their specific sample size is reduced due to the possibility of rejecting the null hypothesis or if it is sufficiently clear that the expected effects are not present and continuing data collection is a waste of resources. Although less attention is typically given to stopping the study for futility before reaching the maximum number of trials, this is not a trivial matter given that resources are limited and collecting more observations might cost more time, money, and effort.

2.2 Methods

The purpose of the experiments conducted in this chapter was to illustrate the application of sequential testing in executive function research. A retrospective analysis of four cognitive tasks was performed to compare the sequential procedure to traditional methods. The evaluation was carried out in timing and accuracy tasks, which measured reaction time and accuracy of responses, respectively. In particular, the sequential testing model of assessment was implemented on Countermanding and Numerical Stroop for the timing tasks, and PASAT+ and

Cancellation for the accuracy tasks. The tasks are described in more detail below.

Regarding the data collection, young adults recruited from the University of California, Irvine and the University of California, Riverside completed mobile-based tests in the app "Recollect the Study", which can be downloaded on the App Store. Recollect the Study has been used as a platform to gather data for other behavioral studies, such as "UCancellation: A new mobile measure of selective attention and concentration" (Pahor et al., 2022). The Recollect app contains a test battery that was designed to measure executive functioning. 18 college students between 18 and 22 years old completed the test battery on touchscreen devices. The tests in the tablet-based battery included the non-threshold-seeking tasks of interest, among other assessments. Since each one of the young adults from the cohort took the test battery ten times, the total number of sessions for each task was 180. The data from these sessions was used to evaluate sequential testing in Countermanding, Numerical Stroop and PASAT+. For the Cancellation task, however, a different dataset consisting of 460 test sessions from college students was preferred for proof of concept. This alternative study was more suitable simply because most of the Cancellation files in the dataset used for the rest of the tasks had an insufficient number of trials to perform a sequential analysis. In both datasets, several data files that contained incomplete sessions or extreme reaction time values (outliers) were removed from the analysis. The final number of sessions for each specific task will be specified in the results section.

The Countermanding task estimates the executive function of inhibitory control or the ability to control the execution of a response by measuring response inhibition latency (Morein-Zamir et al., 2004). The Countermanding version that was used to collect the data was a hybrid of Simon and Spatial Stroop tasks (Davidson et al., 2006; Diamond, 2013). In this test, the participant was instructed to tap on one of two buttons in response to a visual stimulus. Depending on the color of the stimulus, which appeared on the left or right buttons interchangeably, the participant had to tap on the same side of the screen (congruent trial) or on the opposite side (incongruent trial). A brief practice session was followed by three blocks of trials: the first block contained congruent trials only, the second block contained incongruent trials only, and the third and final block contained a total of 48 trials including both conditions (congruent and incongruent). The sequential test procedure was performed on the third block, which was referred to as the “mixed assessment block”.

The Stroop test is one of the famous tests to estimate inhibitory control by measuring interference and its control (Martínez et al., 2018). The Numerical Stroop task employed in this battery assessed inhibitory control based on responses that are either congruent or incongruent with a mental set. In the test battery, a non-symbolic Stroop animal task with two conflicting dimensions (number and size) was performed by the participants. In each trial, they were simultaneously presented various numbers of elephants and frogs on each side of the screen. Then, they had to indicate which side “had more than the other” as quickly as possible, irrespective of animal size. After a few practice trials, the assessment stage began and 60 trials were presented, with equal counts for congruent and incongruent trials.

PASAT+ is an adaptation from the Paced Auditory Serial Addition Task (Gronwall & Sampson, 1974) that combines elements of both a working memory task and a test of information processing speed to measure sustained attention, flexibility, and calculation ability (Tombaugh, 2006). The original task was developed to assess the effects of acquired brain injury on cognitive functioning, but it has also been validated in healthy individuals (Wiens et al., 1997). This test requires attention and is an extremely sensitive measure of vigilance, since the subject is required to not only attend to the relevant numbers, but also be aware of the ongoing changes on the screen. In this assessment, the participants were asked to add the last two elements from a sequence of numbers that appeared on the center of the screen one after another. It should be noted that each number must be added to the one just before it, not to the answer. The sums are pairwise, and the values are selected from a multiple choice set presented at the bottom of the screen. After the practice stage, the responses to 20 assessment trials were recorded.

Cancellation is a timed test akin to the D2 test, a psychodiagnostic instrument for measuring processing speed, rule compliance, and quality of performance (Brickenkamp & Zillmer, 1998). The Cancellation task in this test battery involves the cognitive domains of sustained and selective attention, inhibitory control, psychomotor speed, visual searching and motor coordination (Brucki & Nitrini, 2008). In the version of the task that the participants took, 8 items were displayed per row, with 3 to 5 targets per row (every 10 rows had exactly 40 targets). The goal was to select as many targets and clear as many rows as possible within the global time limit of 3 minutes and 30 seconds for the assessment block. With this, participants could complete “bonus rows” if the global time limit was not exceeded. The time limit for each row was 6 seconds, with 1 second screen blank interval between rows. For each row, the total number of hits, false alarms (Type I errors), misses (Type II errors) and correct rejections was recorded. Each row was considered a trial that was assigned a binary accuracy score: 0 if there were any errors in the row or 1 if the answer was perfect, i.e. , if the row had no Type I or Type II errors.

It is reasonable to presume that the number of trials used to make a decision concerning difference detection is inversely proportional to the amount of uncertainty in the decision. On a case-by-case basis, experimenters may decide if it is more desirable to prioritize short test duration or low error rate, remaining conscious that there is a trade-off between test length and statistical test accuracy. In this study, the outcome from the traditional fixed testing design served as "ground truth" while the sequential procedure estimation was regarded as the "experimental condition". Type I error rate or α corresponds to the probability of a significant result when the null hypothesis (H0) is true, while β is the probability of a non-significant result when the alternate hypothesis (H1) is true.

For each sequential testing experiment, Type I and Type II error counts were determined by observing whether the decision from the standard testing procedure was in agreement with the conclusion from the sequential design. In other words, a final conclusion from an interim look that does not match the conclusion made when the maximum number of observations has been collected is considered an error. For the sake of brevity, "different" is understood in this context as "fluctuations in executive functioning beyond a particular effect size have been detected" and "same" is equivalent to "no fluctuations detected; executive functioning has not changed". Correct detections can be separated between true positives and true negatives, corresponding to both approaches determining "different" and "same", respectively. On the contrary, a Type I error or a false positive occurred when the sequential strategy decided "different" and the conventional strategy decided "same", whereas a Type II error or false negative occurred when the sequential strategy decided "same" and the conventional strategy decided "different".

For each task, contingency tables were used to determine the error rates and fully evaluate the effectiveness of sequential testing with respect to the control procedure. Specificity was calculated as the number of true negatives divided by the total number of false positives and true negatives. Positive predicted value or sensitivity was computed as the number of true positives divided by the total number of true positives and false positives. This metric was used to specify what proportion of positive identifications was actually correct. Negative predictive value was calculated as the number of true negatives divided by the total number of true negatives and false negatives. True positive rate, also known as sensitivity, was computed as the number of true positives divided by the total number of true positives and false negatives. Sensitivity was calculated to determine what proportion of actual positives was identified correctly. Finally, the false positive rate was computed as the number of false positives divided by the total number of false positives and true negatives.

It is important to note that performing multiple tests without correcting the significance threshold, or α level, is associated with an increase in the false positive rate. In fact, inflated error rates due to optional stopping without adjusting the α level is an important problem in the reproducibility crisis. “P-hacking” takes place when researchers violate the preset Type I error probability by rerunning analyses when a statistically significant effect is desired but not found (Head et al., 2015). Therefore, making the α level more stringent when multiple statistical tests are performed is a necessary step to prevent a potential inflation of Type I errors. In the experiments below, α was corrected with the Bonferroni-Holm procedure, but it should be noted that different types of corrections exist. Bonferroni-Holm procedure was found to be the most widely recommended way to reduce the apparent significance of effects from multiple looks (Giacalone et al., 2018). Thus, the rejection criteria were adjusted for each hypothesis test based on the Bonferroni-Holm formula: $\alpha_i = \alpha/(m - i + 1)$ (Holm, 1979). Accordingly, the α

corrected values (α_i) were different for each Mann-Whitney U test or binomial test, depending on three factors: α or the overall Type I error rate, m or the total number of interim analyses, and i or the current hypothesis test index, starting from i = 1.

While guaranteeing low error probabilities, the justification of sequential testing becomes evident when one considers its main advantage: a substantial reduction in test duration. Again, in the case that a large effect is present, a shorter test session with fewer trials could suffice to draw an inference about detectable differences in executive functions. In addition, due to the potential reduction in sample size, all the trials that have not been used on a given session could be invested in additional test procedures or testing sessions, enabling more extensive or more frequent testing for the same time commitment in the long run. In this study, the number of trials that were saved in each test session with sequential analyses was computed to illustrate the efficiency gains of this method. Some sessions required fewer trials due to significance while others due to futility. The futility threshold was fixed for all interim analyses at a particular effect size. The effect sizes were computed with “Cohen's d” and “Cohen’s h2” for reaction time data and accuracy data, respectively. For Countermanding and Numerical Stroop, Cohen's d, one of the most commonly used measures of effect size, was calculated as the mean difference in reaction time divided by the pooled standard deviation (Funder & Ozer, 2019). On the other hand, for the PASAT+ and Cancellation tasks, the accuracy measurements only had two possible values (zero or one), so Cohen’s h2, which is a variation of Cohen’s h that is better suited for binomial tests, was used to measure the distance between the proportions from each independent group (Cohen, 1988).

2.3 Results

The sequential design experiments can be separated into two main categories: models that measure reaction time and models that measure task accuracy. First, the results from timing models reflecting distributions of response times will be presented, and subsequently, the findings from models reflecting derived measures of accuracy will be addressed. For the different group sequential designs, the “cost” of sequential testing is measured by recording the number of Type I and Type II errors, while the “benefit” is indicated by the trial savings relative to the maximum number of pre-planned trials.

2.3.1 Timing Models

Accuracy is typically very high in the Countermanding and Numerical Stroop tasks; thus, the main dependent measures are the mean reaction times for correct congruent and incongruent trials. Performance in these timing tasks can be represented with models of reaction times on a trial-by-trial basis that follow a right-skewed distribution. In other words, the time a subject takes to complete each trial can be visualized in the particular right-skewed distribution from their responses to the cognitive test. The Mann-Whitney U test is a nonparametric alternative to a t test that makes no assumptions about the distribution of the data (Hart, 2001), which might not necessarily be normally distributed. In the group sequential design that was implemented here to test for differences in executive function performance, multiple Mann-Whitney U tests were performed for every session that was compared to a baseline session, such that each Mann-Whitney U test was carried out after each additional batch of samples.

For both timing tasks, the maximum number of sequential Mann-Whitney U tests used was 6, but in practice, many sessions required fewer tests to reach the significance or futility boundaries. As mentioned earlier, the overall α level in a sequential design differs from the α level at each look. For a desired Type I error rate of 0.05, the corresponding α corrected values for each of the six interim analyses were 0.008, 0.010, 0.013, 0.017, 0.025, 0.050. It can be noticed that when the final looks occurred, the α values got closer and closer to the uncorrected alpha level. For Countermanding, the starting number of observations was 8, and after each Mann-Whitney U test, data was added to the sequence in batches of 8, up to a maximum of 48 total trials. For Numerical Stroop, the first analysis took place with 10 trials only, and subsequent interim analyses were performed after incorporating 10 additional trials at a time, up to a maximum of 60 total trials.

Effect sizes play an important role in statistical tests, as they quantify the magnitude of an effect that emerges from the sampled data (Schäfer & Schwarz, 2019). Naturally, the larger the magnitude of the difference between groups, the lower the number of samples required to detect it. Thus, it is anticipated that the individuals that would benefit the most from the sequential design are those that show larger differences in performance between sessions. If their performance is almost identical, they would also take very short tests because the effect size is too small at early inspections to be detectable, and it is not worth wasting more trials and testing time. Effect sizes for timing data can be computed with Cohen's d. Cohen originally classified d values of less than 0.2 to be small, and d values greater than 0.8 to be large (Sullivan & Feinn, 2012). In general, very small effect sizes will correspond to non-significant p values in the Mann-Whitney U tests, indicating that there are no detectable differences in executive functioning. By contrast, large effect sizes lead to significant p values, and the conclusion will be that significant differences in task performances are detected.

In sequential testing for timing models, data collection was stopped for futility by rejecting the presence of an effect of interest, but an appropriate futility threshold that does not excessively raise error rates must be established. For the timing experiments, data collection was stopped early if d was less than 0.05 at any of the sequential tests, and the alternative hypothesis was rejected. In sequential designs, stopping for futility because the final result is unlikely to be a significant result has been coined as "stochastic curtailment" (Lakens et al., 2021). This practice saves testing time due to the low probability of observing a significant effect within the predetermined maximum sample size. Only in extreme samples with large variation would a Cohen's d value smaller than 0.05 at an interim analysis eventually reach statistically significant results. With lower data requirements, futility implied the following conclusion: "same executive function state: no detectable changes".

For the Countermanding task, 166 sessions were compared pairwise to each other, considering one session the "Day 0" baseline that was compared to the other session in search of changes relative to the reference day, which was evaluated using the entire trial budget. Summary statistics were computed from the values in the contingency table shown in Figure 2.1, which indicates the error counts for this task. Specificity was 3108 / (345 + 3108) = 0.90, true positive rate (sensitivity) was 9792 / (9792 + 616) = 0.94, false positive rate was 345 / (345 + 3108) = 0.09, positive predictive value (precision) was 9792 / (9792 + 345) = 0.97 and lastly, negative predictive value was 3108 / (616 + 3108) = 0.835. Figure 2.1 also shows a pie chart of the trial savings for the Countermanding task. Notably, 42.1 % of the sessions used just the initial 8 trials, which was the minimum number of observations.

For the sequential design, the average number of trials was 20, and only 16.3 % of the sessions used 48 trials, which was the maximum number of pre-planned observations.

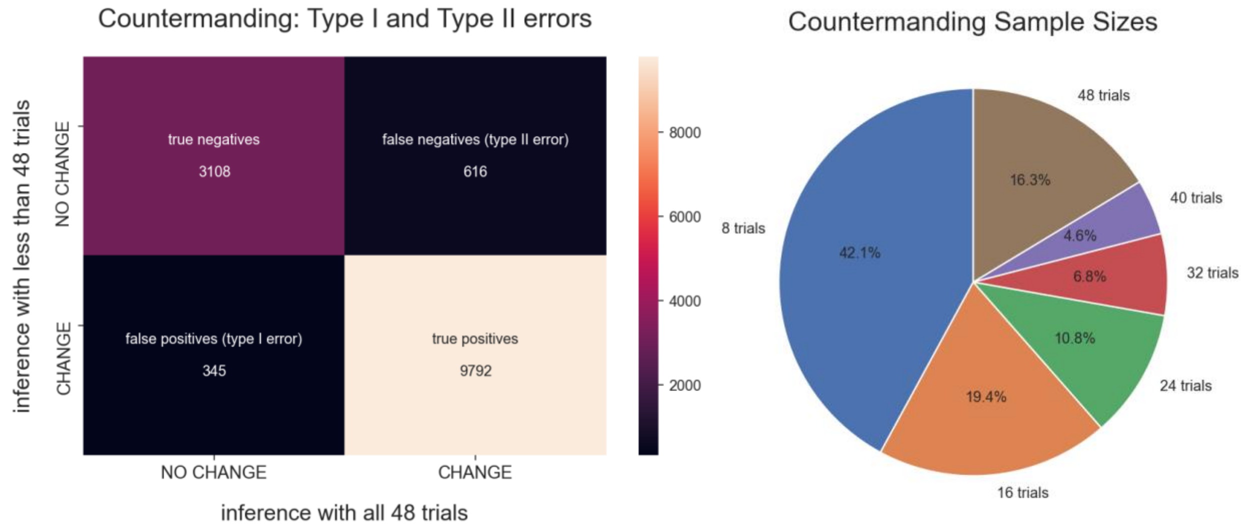

Figure 2.1: Countermanding error rates and trial savings with a significance level of α = 0.05 and futility of d ≤ 0.05.

A sequential design demonstrated more efficiency than a fixed design in the Numerical Stroop task, too. Figure 2.2 shows a contingency table with the count of Type I and Type II errors for a total number of 5886 comparisons of the 108 sessions of the task against each other. A pie chart with a detailed breakdown of the number of trials used for each session is also included in Figure 2.2. Specificity was 1265 / (214 + 1265) = 0.86, true positive rate (sensitivity) was 3995 / (3995 + 412) = 0.91, false positive rate was 214 / (214 + 1265) = 0.15, positive predictive value (precision) was 3995 / (3995 + 214) = 0.949 and negative predictive value was 1265 / (412 + 1265) = 0.754. The trial savings were evident, given that 88.2 % of the sessions were shorter using sequential tests and that more than fifty percent of the sessions only used 10 or 20 trials. The average number of trials was 23 for the sequential test design, which was considerably smaller than 60, the number of pre-planned trials used in the fixed design.

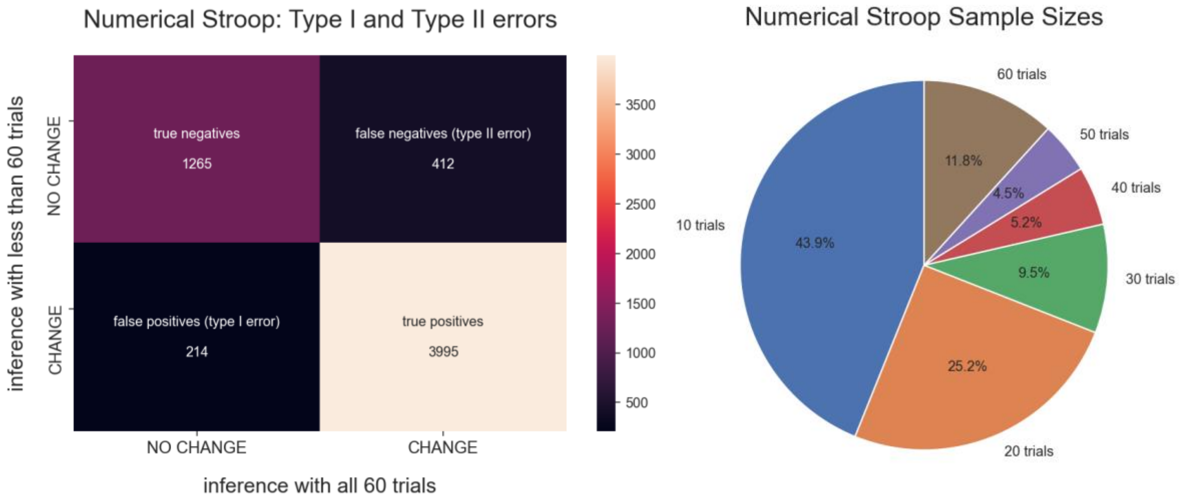

Figure 2.2: Numerical Stroop error rates and trial saving with a significance level of α = 0.05 and futility of d ≤ 0.05.

2.3.2 Accuracy Models

In this subsection, group sequential designs are analyzed using a different statistical test comparing accuracies instead of reaction times. PASAT+ and Cancellation are tests whose outputs are typically quantified as a proportion of correct trials. Performance in such accuracy tasks can be represented by modeling the accuracy across all the trials with a binomial distribution. As described earlier, the approach selected to reduce testing time of accuracy tasks was sequential binomial tests. In these two tasks, each trial had two possible outcomes (i.e. success/failure), so the two-sided binomial tests were used to determine whether the observed test accuracies from a new testing session differed from what was expected (accuracy from a baseline session). The input argument for each interim binomial test was the cumulative number of successes and the total number of independent trials presented, while the output was the corresponding p value. For PASAT+, the starting number of observations was 4, and after each binomial test, data was added to the sequence in batches of 4, up to a maximum of 20 total trials. For Cancellation, the first analysis took place with 10 trials, and subsequent interim analyses were performed after incorporating 10 additional trials at a time, up to a maximum of 40 total trials.

For both accuracy tasks, the overall level of significance was set at an α value of 0.05, which is the exact same Type I error rate that was set for the timing tasks above. For PASAT+ the corrected alpha levels for each of the five sequential binomial tests were 0.010, 0.013, 0.017, 0.025, 0.050. On the other hand, the α thresholds for each of the four sequential tests carried out in Cancellation were 0.013, 0.017, 0.025, 0.050. Regarding futility, a threshold was set for each task using a minimum effect size in terms of Cohen's h2 because the data for the accuracy tasks was binary (ones and zeros), rather than continuous (reaction times). For PASAT+ and Cancellation, the futility boundary was set at a Cohen's h2 value of 0.1. Effect size values below the futility threshold were considered to be negligible differences in performance for this study. One should not forget that higher futility boundary values could have been set to reduce even more the number of trials, but at the cost of higher amounts of false negatives or Type II errors. Again, the cost and benefits of the sequential designs are reflected in contingency tables showing Type I and Type II errors, and pie charts illustrating how much test length was reduced for each task.

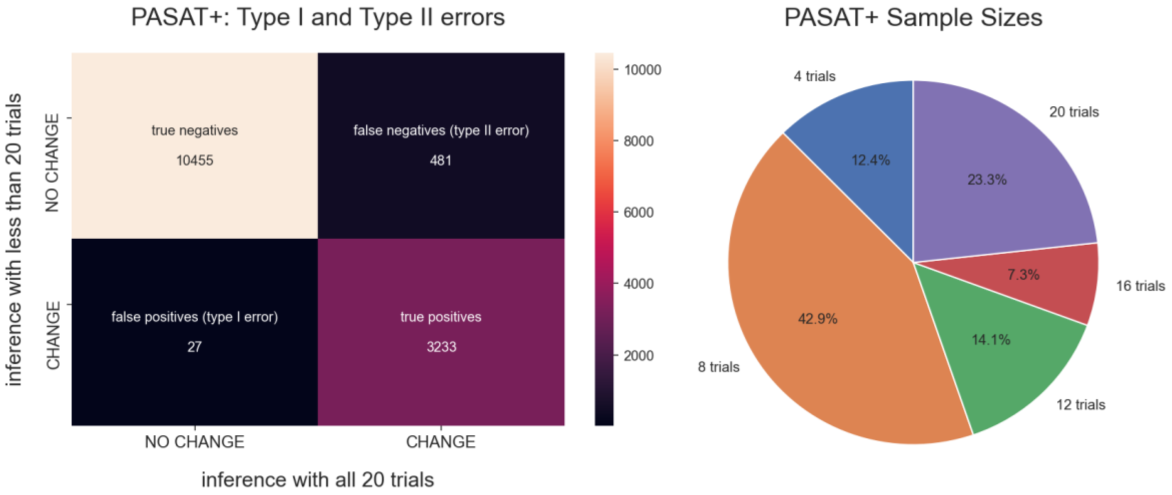

Figure 2.3: PASAT+ error rates and trial saving with a significance level of α = 0.05 and futility of h2 ≤ 0.1.

Figure 2.3 was obtained by comparing 170 sessions of PASAT+ against each other in search of changes in executive functions. Using sequential binomial tests, specificity was 10455 / (27 + 10455) = 0.997, true positive rate (sensitivity) was 3233 / (3233 + 481) = 0.870, false positive rate was 27 / (27 + 10455) = 0.00258, positive predictive value (precision) was 3233 / (3233 + 27) = 0.992 and lastly, negative predictive value was 10455 / (481 + 10455) = 0.956. The average number of trials was 11, and 76.7 % of the sessions were shorter than the traditional test. Again, the sequential tests could have adopted less conservative stopping criteria, and further improved the time savings for this task. For instance, the large positive predictive value indicates that the false positive rate is very low, as shown in the contingency table. Perhaps, a slightly greater value than 0.05 could have been chosen for the α level in PASAT+.

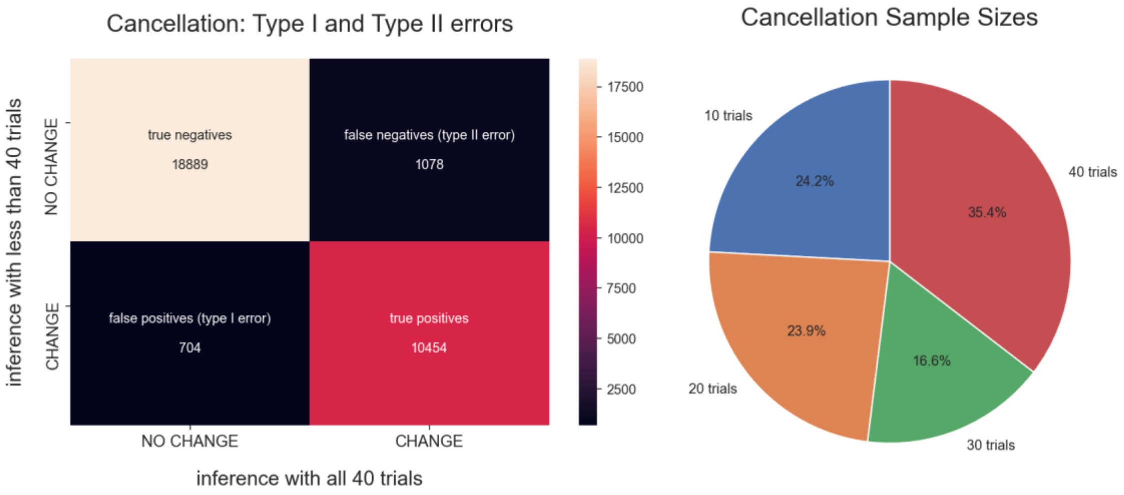

Figure 2.4: Cancellation error rates and trial saving with a significance level of α = 0.05 and futility of h2 ≤ 0.1.

Finally, Figure 2.4 reflects the difference detection summary for Cancellation, where 249 sessions were compared against each other. Specificity was 18889 / (704 + 18889) = 0.964, true positive rate (sensitivity) was 10454 / (10454 + 1078) = 0.907, false positive rate was 704 / (704 + 18889) = 0.036, positive predictive value (precision) was 10454 / (10454 + 704) = 0.937 and lastly, negative predictive value was 18889 / (1078 + 18889) = 0.946. The average number of trials was 26, and 65.6 % of the sessions were shorter than the traditional test. The trial savings were relevant but perhaps not as substantial as they were for the other tasks above. This could be attributed to the fact that traditional output metrics for Cancellation rely solely on accuracy, but in this task the speed of the response also plays an important role. Maybe the reason why more trials are required to detect significant changes in performance in Cancellation is that reaction time differences are not being assessed in the sequential tests, which only test for differences in the number of perfect rows. An even more relevant aspect that also needs to be considered is that the trade off between trial savings and error rates is affected by the effect sizes that an experimenter is trying to measure. As a consequence, the experimental outcomes in the contingency tables and pie charts for all the tasks above will vary depending on the minimum effect size of interest and the conservativeness of the error rate threshold. This study established the α significance level at 0.05 and the minimum meaningful effect size at 0.05 for Cohen’s d and 0.1 for Cohen’s h2. These values, which were empirically determined, yielded satisfactory error rates and trial savings, but other upper and lower stopping boundary combinations could be attempted too.

2.4 Discussion

A frequent limitation of current behavioral test procedures is assessment length, which is one of the most relevant factors responsible for respondent burden (Kleinert et al., 2021). Unfortunately, fixed testing designs are standardized and can take a long time to complete, making it inconvenient to get quick diagnoses of changes in executive functions. Because of their block design, conventional tests must proceed to completion and use the maximum sample budget before the analysis stage. Current methods for drawing inferences about latent constructs would derive benefit from the ability to speed up the search for differences in executive functions in cases where effect sizes happen to be relatively large. To this end, more efficient data collection strategies that decrease test duration, and thus mitigate respondent burden, need to be developed. In threshold-seeking tasks, such as Corsi span tasks, classical active learning with modulation of task difficulty typically leads to a significant reduction in data requirements. In fact, in the next chapter, it will be demonstrated that active sampling techniques can produce fast and accurate psychometric function estimation.

At the individual task level, a different approach is needed for improving efficiency in assessments designed without independent variables, such as task difficulty or stimulus intensity. Essentially, many non-threshold-seeking tasks consist of repeating the same task item multiple times or systematically delivering the same counterbalanced task items for each individual. In this setting, because all the trials are inherently identical and there is no independent variable to manipulate, specific selection of testing items is not applicable. Therefore, the number of total task items to deliver becomes the only aspect that the experimenter can adjust. Effectively, the number of trials is directly proportional to task duration, which is the main feature that can be modified in non-threshold-seeking tests through active testing. Batch sequential testing was proposed in this chapter as a useful approach to actively select the number of trials and detect individual changes in executive functions in shorter periods of time. Although underutilized in most scientific research domains, sequential testing is a well-established procedure that can be widely adopted for active testing of executive functions. Eventually, the ultimate solution is expected to consist of multidimensional latent variable models that exploit the relationships between results of repetitive tasks.

To provide evidence of the efficiency gains of sequential cognitive testing, quality assessments of the models for non-threshold-seeking tests have been reported. Broadly, the evaluation of group sequential testing in several timing and accuracy tasks aimed to bring this existing methodology to the attention of researchers in the behavioral science field. The objective of the adopted strategy was to interrupt the data collection process as soon as the number of observations gathered was sufficient to detect significant changes in task performance or to conclude that there were no detectable changes because the measured effect was too small to be resolved even with the maximum number of planned task items. Performance of sequential testing was satisfactory since the error rates were reasonably low and efficiency gains were achieved. On average, sequential testing required smaller sample sizes than traditional methods, as illustrated in the pie charts above. Despite using fewer test items to make a determination of “same” or “different”, the Type I and Type II error counts in the contingency tables were relatively low, suggesting that researchers should contemplate sequential procedures as a preferable alternative for saving time, effort and other resources in the long run. Overall, this chapter revealed the potential utility of sequential testing and emphasized a long-held view that the data requirements in traditional cognitive tests are excessive.

In general, performing multiple statistical tests to sequentially monitor the data as it accrues requires careful pre-planning. This could perhaps be argued to be one of the disadvantages of sequential testing compared to non-sequential approaches. A few of the decisions that researchers need to make beforehand are what are appropriate effect sizes for difference and futility for the task at hand, what the maximum sample size will be, how much time will be dedicated for interim analyses, at what points will the interim analyses be performed, and so on. However, one should also be conscious that there are several software tools available to design and analyze sequential batches of trials more easily. An important advantage to emphasize is that this procedure is based on traditional frequentist statistics, which are methods that most researchers are familiar with and that generally involve low computational complexity. Rather than using complex machine learning tools, this strategy is simply founded on well-known statistical methods, such as U tests and binomial tests, and thus can be viewed as a fallback when more intricate techniques are not feasible.

Sequential testing results in many benefits to cognitive and perceptual evaluations. The flexibility of this approach allows experimenters to make modifications to the test characteristics to each unique task and target population. Statistical test accuracy and test length can be traded off by adjusting the effect size of interest, the expected false positive rate or the expected power. Furthermore, the interim capacity of dynamically adjusting each test individually can impact the expansion of sequential executive function testing to a wide variety of contexts, regardless of hardware capacity. Ultimately, another long-term benefit of this approach could be increased engagement and participation in regular behavioral assessments that would lead to continuous monitoring of executive functions across time. Similar difference detection frameworks that enable substantial test length reductions could be implemented in future studies in a variety of cognitive domains beyond the tasks addressed in this thesis.

Chapter 3: Individual Psychometric Probabilistic Model

In the field of psychometrics, much effort has been spent on unidimensional estimation tasks over many decades, but surprisingly little work has gone into estimating multidimensional models, using novel estimators besides linear regression. In higher dimensional spaces, predicting individual task performance person by person in an effective and efficient manner is a formidable challenge. Combining behavioral measurements across subjects from a cohort reduces noise as well as testing time commitment from each person, but it can be problematic in heterogeneous populations where observations vary considerably from individual to individual. Low proportion of particular subgroups in the population comprising the standardization sample inevitably introduces systematic errors in detriment of underrepresented groups. In other words, referencing individuals against one another and norming raw measurements is a major source of testing bias (Reynolds & Suzuki, 2012). To date, rigorous estimation procedures for complex multidimensional problems are neither standardized nor well-established, as they would inescapably entail burdening participants with large numbers of often tedious trials.

The threshold-seeking psychometric test design for the working memory task described in this chapter is ahead of the repetitive testing addressed above. In chapter 2, sequential testing was essentially one of the very few alternatives available to improve efficiency in non-threshold-seeking tasks because they lack independent variables. Threshold-seeking span tasks, however, are modeled by a psychometric function that depends on an explicit input variable. To improve upon span tasks and similar cognitive tests, a number of approaches could potentially be developed in response to the lack of estimation procedures in higher dimensions. Eventually, it would be of particular interest to create more thorough models that allow for multiple forms of measurement to contribute towards a better understanding of cognitive and perceptual underlying constructs such as latent variable models.

One of the main objectives of the research line described in this chapter is to take a step forward in building a multidimensional estimator capable of incorporating a wide variety of prior beliefs and co-estimation procedures into an active learning process. It is noteworthy that this study effort promotes more equitable evaluations of latent constructs by developing estimation techniques that do not rely exclusively on variability between individuals. Advantageously, modern machine learning approaches have indeed become a major enabler for unified models capable of reflecting complex trends across multiple variables. Pursuit of this approach can provide a basis for a simultaneous speed up in estimation convergence with fewer observations and a richer summary of each individual's task performance.

The key development effort that is presented in this chapter is the establishment of an improved Gaussian Process (GP) estimator for the 4-parameter psychometric function model. The technical details of GP-based modeling will be explained more fully below, but it is worth underscoring that a major contribution of this work was the generalization of a novel binomial likelihood function. Previously, only a Bernoulli likelihood had been implemented in a GP framework (Chen, 2020). Currently, the initial testing has been undertaken in individual psychometric functions yielded from accuracy scores of a span task delivered to a young adult population. More specifically, this probabilistic machine learning classifier has been validated in a spatial working memory task quantifying a person's ability to recall a sequence of spatial locations. This step has been fundamental to confirm the desired behavior of the estimator when it comes to constructing complex individual models of executive functions in one dimension. Ultimately, the work described in this chapter paves the way to expand the estimator into new dimensions in order to incorporate other informative predictors.

3.1 Psychometric Functions and Gaussian Process Framework

Psychometric functions are probabilistic predictors of behavior that represent the relationship between a certain parameter of a perceptual or cognitive phenomenon and a subject's performance on a task (Wichmann & Hill, 2001). Hence, they represent a mathematical model of response probability as a function of a stimulus feature that is useful for inference of psychological constructs (Gold & Ding, 2013). This analytic function can be fully described by four parameters: the threshold, the spread and two additional parameters that define the upper and the lower asymptotes (Treutwein & Strasburger, 1999). Much emphasis has been placed on estimating the value at which a subject achieves some arbitrary proportion of correct detections: the threshold at a predetermined percent correct. Most standard procedures, such as adaptive staircase, have focused on determining performance threshold directly, without estimating the entire psychometric function (Shen, 2013). Nonetheless, it is worth highlighting that estimation of all four parameters, beyond just the threshold, is crucial for a more comprehensive evaluation of underlying brain processes. Furthermore, generating a probabilistic model of the full function enables efficient active model selection (Gardner et al., 2015; Larsen et al., 2021).

Common psychometric function models include the Gaussian and Weibull cumulative distribution functions (Żychaluk & Foster, 2009). The cumulative Gaussian distribution function is given by $\Phi\{x; \alpha, \beta\}$, where α corresponds to the aforementioned psychometric threshold and β, the spread or inverse of the slope of the curve at threshold, quantifies the transition from task success to failure, thus reflecting internal task process noise (Buss et al., 2006; Strasburger, 2001). Due to chance or task design, poor fits to the 2-parameter model may arise when responses to stimuli at the low and high asymptotes of the psychometric function are not in agreement with model predictions. However, estimates of the threshold parameter and especially the spread parameter can be quite inaccurate if the asymptotes of the sigmoid function are not estimated concurrently (Wichmann & Hill, 2001). In consequence, building accurate fits that do not make strong assumptions about the shape of the curve requires a more general model that is allowed to be shifted and/or scaled along the ordinate axis. Accordingly, the cumulative Gaussian distribution function is augmented and leads to a new model that is analytically expressed as $\Psi(x) = \gamma + (1 - \gamma - \lambda)\ \Phi(x)$ (Kingdom & Prins, 2010; Wichmann & Hill, 2001). Of note is that this equation includes not only psychometric threshold (α) and spread (β), but also lapses (λ) and guesses (γ). The two additional parameters, λ and γ, have a crucial impact on the extremes of the psychometric function because they capture response behavior at low-probability events. While the lapse parameter accounts for deviations from perfect performance at easy tasks where correct responses are almost always expected, the guess parameter contemplates the possibility that subjects may guess the correct answer for a difficult task with nonzero probability. Even though λ and γ are generally considered nuisance parameters unrelated to the construct of interest, they still play a significant role in constructing a complete model that exploits the information yielded by the entire sigmoidal psychometric function.

Logistic regression, a classifier version of linear regression, is the gold-standard estimator of a 4-parameter psychometric sigmoid representing probabilities of task performance (García-Pérez & Alcalá-Quintana, 2005; Kingdom & Prins, 2010; Yssaad-Fesselier & Knoblauch, 2006). In more recent works, another tool that has been employed to model a psychometric function is a Gaussian Process (GP). GPs are distributions over functions that serve as useful models for probabilistic inference about underlying constructs. As probabilistic models, they are taken as a valuable metric to prediction uncertainties in our estimations. It is also worth noting that parametric logistic regression has been shown to be functionally equivalent to a GP classifier in the one-dimensional case (Song et al., 2018). A significant advantage of using a GP model is that it is defined only by the input data due to its non-parametric nature. Because estimations are derived from the observations and not a set of fixed parameters, this powerful model provides unprecedented flexibility for estimating a wide variety of functions. Because of the scalable nature of this modeling procedure, multiple input dimensions can be included with practical amounts of data in a real setting. Further, a GP combines prior beliefs with new data to generate a posterior belief about the latent function, granting more efficient estimation and more powerful inference chains in higher dimensions (Rasmussen & Williams, 2006; Williams, 1998). By virtue of active learning algorithms, this Bayesian method can place its allowance of trials at the most informative input values for each individual, and therefore cut down massive data requirements dramatically.

The 4-parameter model described above, $\Psi(\mathrm{x}) = \gamma + (1 - \gamma - \lambda)\ \Phi(\mathrm{x})$, acts as a linking function of the latent function. This means that $\Psi(\mathrm{x})$ compresses the output range of the GP from the range of all real numbers ($\mathrm{f}(\mathrm{x}) \in \mathbb{R}$) to the range of probability values ($\mathrm{f}(\mathrm{x}) \in [0,1]$). One caveat of the 4-parameter model is that, in some cases, the nuisance parameters that this linking function integrates could affect the shape of the sigmoid more than desired. Fortunately, owing to the Bayesian nature of this framework, investigators can constrain the model by incorporating their domain-specific knowledge into it through deliberate specification of a prior distribution for each of the estimated parameters (Treutwein & Strasburger, 1999). For this reason, a logical solution to prevent mislabeled data points from having a detrimental effect on accurate model fitting is to establish a probability prior on γ and λ. In this work, a beta prior defined on the interval [0, 1] and parameterized by two positive shape parameters, a = 2 and b = 50, was incorporated into the model. The mode of this prior distribution was 0.02, indicating a general expectation of a 2% probability of lapses or guesses. Simply put, the implementation of this beta prior ensured that the effect of the guessing and lapsing rates on the psychometric fits was only modest.

In this GP framework, there are two relevant elements that must be combined to the 4-parameter linking function, namely, a likelihood and a kernel. The likelihood function represents the information gained from new data generated from the working memory construct, and it is equivalent to the sampling distribution probability function $\mathrm{p}(\mathrm{y}|\theta)$ for fixed observations y, given the model parameters θ. Although the GP is a non-parametric model, the components of the GP, such as the kernel function, may themselves have parameters. These are referred to as hyperparameters, and their correct adjustment exerts great influence over the predictive distribution of the GP (Barbour et al., 2019). Typically, a Bernoulli, Gaussian or binomial likelihood is a valid choice of likelihood function compatible with a GP describing a latent function. In this particular unidimensional working memory application, the GP estimates performance based on observations at repeated input values in the task difficulty space, so a binomial likelihood is best suited to describe the latent function. On top of this likelihood, a kernel function must be selected to encode information about the shape and smoothness of the functions drawn from the GP. A linear kernel as a function of sequence length is useful to manifest the property of monotonicity of the function $\Psi(x)$. Naturally, task performance in the spatial working memory test is expected to be monotonic with sequence length because the probability of successfully completing longer (harder) trials is expected to systematically diminish. The posterior belief about $\Psi(x)$ will tend toward 1 for the easiest trials and then, as task difficulty increases, it will gradually descend toward 0. It is important to point out that the covariance function of the linear kernel captures any deviation from the central tendency of the latent function, and that the mean function depends on the prior mean and the posterior covariance. This implies that an arbitrary value can be assigned to the prior mean function, which was set to a constant value of 0 in this framework. Altogether, the linear kernel was combined with the 4-parameter linking function, alongside the binomial likelihood, to build our GP probabilistic classifier.

3.2 Model Training

In Bayesian statistics, the basis for drawing inferences is the posterior distribution, which is a combination of the prior knowledge and the observed evidence (van de Schoot et al., 2014). Bayes' theorem requires computing the product of the likelihood function and the prior distribution for obtaining the exact form of the posterior distribution (Etz, 2018). A frequently faced problem is that inference in probabilistic models might be intractable, and thus, calculating the posterior distribution poses significant computational challenges. To tackle this issue, posterior approximation inference algorithms have been developed to provide approximate solutions to the inference problem (Park & Haran, 2018). In the case of GP classification, the categorical nature of binary responses leads to a non-Gaussian posterior that can be approximated with methods such as variational inference. This particular technique casts inference as an optimization problem and aims to maximize the log marginal likelihood log $p(y|X, \theta)$ in order to estimate the hyperparameters from the data. Following this inference method, the best approximation possible is accomplished by minimizing the Kullback-Leibler divergence between the variational inducing distribution (variational GP) and the prior inducing distribution (true GP posterior) (Matthews et al., 2016).

Sparse GPs are a main avenue for addressing the challenge of high time and space complexities in large-scale GP regression. This approach has become an attractive strategy to reduce computational complexity and obtain faster convergence for GP approximation using a set of inducing points (Cheng & Boots, 2017). This subset of fictitious data points is judiciously selected from the initial data points, but it should be acknowledged that determining which specific points to retain from the input domain for optimal training of the GP is not a straightforward matter. Since the number of inducing points is directly proportional to the expressiveness of the full GP, there is a trade-off between the generalizability of the GP approximation and the number of inducing points (Quiñonero-Candela & Rasmussen, 2005). In this working memory task, the sparse framework was included by setting an inducing point at each of the 9 integer-valued locations that correspond to different levels of task difficulty, or at each data point abscissa, in the case of low data counts.

In the process of building Gaussian process models of psychometric curves, an important matter to consider is the choice of prior beliefs. In this work, uninformative prior beliefs were used for model hyperparameters when fitting GPs to human data. Put another way, all possible hyperparameter values were considered equally likely before the first data point for a given individual was incorporated into the model. The uninformative priors that were used for all human data models had no discernible impact on model accuracy or efficiency. Then, human data served as priors for the simulations to aid in selecting appropriate trials for new subjects in the group, while allowing new data to overrule this information if it degrades model quality. Concretely, prior beliefs for model hyperparameters were given as gamma distribution fits to the hyperparameter distributions observed from GP models from using uninformative priors (Kuss et al., 2005). In this unidimensional setting, informative priors had no discernible impact on ultimate model accuracy or efficiency.

3.3 Working Memory Task

When performing a task or solving a problem, working memory fulfills the role of consciously selecting relevant information, retaining it for short periods of time and then manipulating it to plan and guide behavior (Cortés Pascual et al., 2019). Broadly, the more verbal or visual-spatial elements that must be held in mind, the more difficult the recall task becomes. In this context, a psychometric function is a suitable representation of the working memory construct to visualize the proportion of correct detections as a function of stimulus level. Therefore, for varying numbers of item counts, the probability of successfully remembering a sequence of a certain length is reflected in the psychometric function, which maps the latent construct to the observed responses. A spatial working memory task was selected as a simple unidimensional test case to demonstrate the utility of the GP classifier. A widely used test to measure visuo-spatial working memory is the Corsi block tapping task (Corsi, 1972), which has been traditionally administered with a physical wooden board. The advent of numerous computer-based versions over the recent years, such as eCorsi, allowed for the test to be completed on tablets and other digital devices (Brunetti et al., 2014). A modified version of the standard eCorsi was used in this study, with 9 and 12 spatial locations on the screen, for children and young adults, respectively (Ramani et al., 2020). The discrimination task consists of recalling the exact order of appearance of a sequence of targets at the different locations.

In this simple span task, the GP model samples one integer-valued sequence length at a time. Target sequence length is an independent variable that is commensurate with task difficulty, given that it is harder to remember longer random sequences of items (Towse et al., 1998). The dependent variable, task accuracy, represents the probability of correctly reporting a sequence of elements that was presented to the participant. At each trial, a GP is calculated, and the response is classified as either correct or incorrect. The sigmoid-shaped psychometric curve determines class membership, and, intuitively, the trials around the threshold point are expected to be the most informative for summarizing performance. In this way, the GP model is a probabilistic classifier that helps to determine the boundary above which a person cannot successfully perform the task, allowing one to draw inference about the associated cognitive function.

Familiarization with the task was facilitated in the form of two unscored practice trials that were provided before the assessment stage began. The participants were then granted two lives and were presented the first scored trial at a sequence length of two. Thereafter, sequence length was increased by one if the response was correct. If a trial was recalled incorrectly, sequence length stayed the same for the next trial and would then decrease by two if the response was incorrect after the second attempt at the same sequence length. Every time a participant consecutively answered incorrectly a particular set of items, sequence length was decreased by two and a life was lost. The ascending or descending levels of difficulty follow an adaptive staircase procedure, where a reversal, i.e. a change of direction in stimulus difficulty, is elicited by a change in response. The output metric that was originally used for this task was maximum sequence length, i.e., the largest sequence length that was successfully completed. In place of this traditional measurement of task performance in the adaptive procedure, threshold estimates from the GP model were used in this new analysis to quantify performance.

To validate the GP model, human data was collected from a cohort of 323 college students between 18 and 22 years old that completed the simple span task on touchscreen mobile devices. One participant from the initial cohort was excluded from the analysis due to task incompletion, whereas 17 subjects were removed because their psychometric model fits were unreasonable, most likely because of the small amount of data. The specific exclusion criteria were threshold estimates greater than 15 and/or spread estimates greater than 10. Overall, the total count of individuals that made up the analytical sample was reduced to 305. On a final note, it is important to highlight that even though the data used in this analysis was gathered from young adults, this novel framework is expected to be easily generalizable to other populations because the GP model fits are individualized. That is, the validity of the spatial working memory task could be demonstrated for individuals of different backgrounds and age ranges.

3.4 Experiments and Results

In this spatial working memory task, performance of the GP model was first assessed using real data from the 305 individuals. The proposed GP estimator was used to estimate a 4-parameter psychometric curve for each one of them. Model goodness-of-fit was quantified by root-mean-square error (RMSE) between each model and the experimental data. Secondly, several simulation conditions were used to assess the capability of the generalized binomial likelihood to create unbiased estimates of 4-parameter psychometric functions. Sigmoidal observation models for classification, comparable to the simple span accuracy curves, were established as ground-truth generative models of participant responses. In addition, the GP model learned all the psychometric parameters α, β, γ and λ in order to construct the 4-parameter model Ψ. To determine how reasonable the resulting GP fits were, prior psychometric curves from the population data served as a valuable point of reference to confirm that the estimated parameters fell within the expected ranges.

3.4.1 Model Validation: Real Data Fits

As shown in Figure 3.1A, very good model fits to the data collected from the population were accomplished, since most RMSE values lay close to 0. The chances are that many of the high RMSE values corresponded to models that were not as good at capturing the observed trends due to smaller amounts of data and high variability. Representative example individuals of the population, whose RMSEs are depicted by vertical red lines in Figure

3.1A, were selected for the subsequent simulated data collection procedures. The joint distribution of psychometric thresholds and spreads was used to identify those participants, which resided at the joint 10th, 50th and 90th percentiles of both parameters. Model fit quality was inversely proportional to variability in task performance around threshold. This trend is reflected in Figure 3.1B-D, showing psychometric functions at the 10th, 50th and 90th percentiles of fit quality.

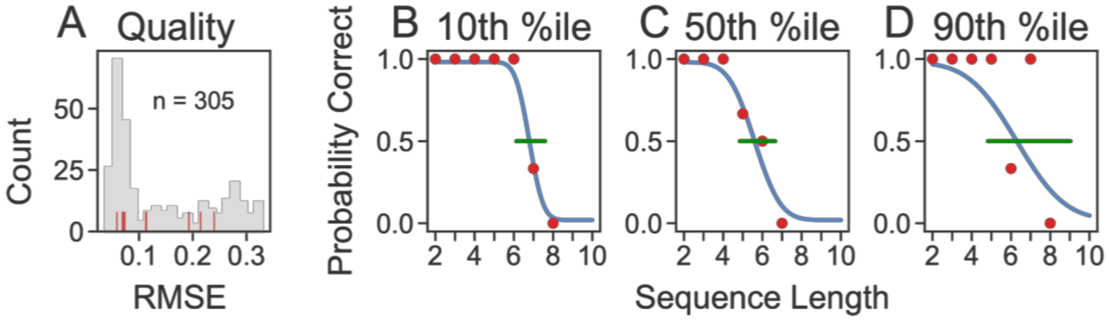

Figure 3.1: A. Histogram showing goodness-of-fit distribution over the cohort using the GP model and data collected using an adaptive staircase design. Vertical red lines indicate 9 individuals selected for more extensive analysis. B-D. Psychometric functions estimating probability correct at the 10th, 50th and 90th percentiles of RMSE values for this population are shown by the blue curves. Red circles indicate the proportion correct values observed at each sequence length. Green lines indicate the 95% credible intervals for the threshold estimates.

For all members of the study population, thresholds and spreads were determined from 4-parameter model fits, as illustrated in Figure 3.2. In conjunction with the trend toward lower RMSE values in Figure 3.1, the considerable variation in parameter values appears to be at least partly attributable to variability in actual working memory across the cohort. In other words, true working memory trends at the population level seem to be captured by the data collection procedure. Furthermore, underpowered psychometric fits for small data sets are another potential source of variability in threshold values. Model goodness-of-fit tends to be highest at spreads between about 0.5 and 1, probably due to the combination of finite resolution of memory load (i.e., task difficulty) and the limited data collected. Individuals at the right of the scatterplot had higher thresholds, indicating greater working memory capacity. The plot symbols toward the top of the scatterplot correspond to subjects with higher spreads, and thus greater internal memory process noise. Most participants cluster between threshold values of 4.5 and 8, and spread values of 0.5 and 3. The representative models that were further analyzed in the simulations are the 9 red crosses of the cohort members closest to joint threshold and spread percentiles of [10, 50, 90].

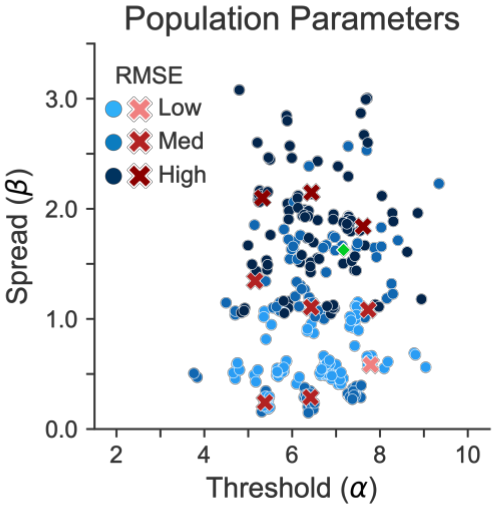

Figure 3.2: Scatterplot of estimated threshold and spread values showing that most results are clustered in a compact joint domain. One individual evaluated with 20 times the average data is indicated by a green diamond. Nine individuals indicated by red crosses were selected for more extensive analysis throughout the chapter. The shading of each data point indicates the RMSEs of the model.

The fundamental premise of this study is that working memory performance can be adequately captured using properties of psychometric functions, such as psychometric thresholds. Thresholds are a proxy for the memory load at which individuals would be equally likely to correctly versus incorrectly recall a target sequence. A traditional metric that has been used to capture internal memory processes is maximum sequence length. This common method of quantifying task performance is defined as the largest sequence length successfully recalled during the adaptive testing procedure (Conway et al., 2005). If both maximum sequence length and psychometric threshold capture similar trends over the population that took the simple span test, a strong correlation between them is expected. To test this hypothesis, a comparison between maximum successful sequence length recalled and estimated psychometric thresholds was carried out. The linear relationship determined in the study population between these two performance metrics is plotted in Figure 3.3. The inset at the bottom of the figure shows the distribution of maximum sequence length values. The histogram confirmed that the testing procedure as designed can successfully bracket the working memory performance of young adults. These results were in agreement with our assumption that both performance metrics are reflective of similar brain processes. Overall correspondence was high with a coefficient of determination of 0.812.

Maximum sequence length, which was about one unit lower than threshold on average, was mapped to a compact subrange of thresholds. Thresholds are real-valued, but maximum sequence lengths can take on integer values only, yielding lower resolution. On the condition that both performance metrics are equally reliable, the higher resolution thresholds represent a potential advantage of fitting psychometric functions to data of this sort. With higher resolutions one has the ability to more finely determine whether the best models for two different data sets are the same or different. This benefit is key for the long-term goal of this project: quantification of the dynamic performance of working memory and related domains.

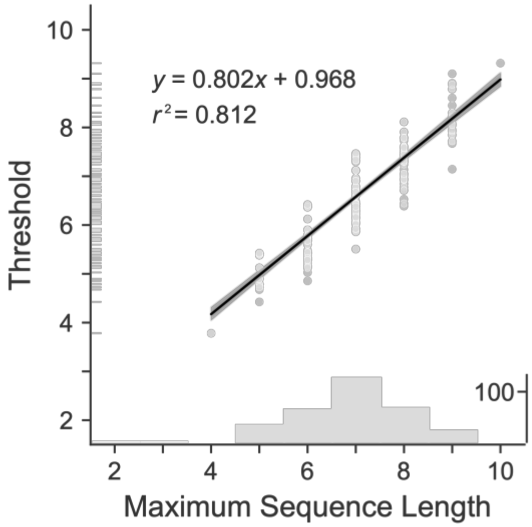

Figure 3.3: Scatterplot of psychometric threshold versus maximum sequence length for the study population shows a linear relationship with a coefficient of determination of 0.812. Estimated thresholds are on average about 1 unit less than maximum sequence lengths. Margins show details of population distribution of each variable.

For the purpose of evaluating the accuracy of the GP psychometric function estimator, it is useful to extensively analyze and visualize fits to a wide variety of task performance settings. Modeled psychometric functions to task data from nine representative individuals that cover most of the range of goodness-of-fit values across this study population are depicted in Figure 3.4. From the perspective of the modeling process, they were selected to reflect joint variation in threshold and spread across the study population, encompassing the diversity seen in the full cohort.

Observation variability, psychometric spread, and model goodness-of-fit appear to be correlated in Figure 3.4 with posterior beliefs about interval estimates of the threshold values. This result is a key rationale for developing GP models of executive functions that, with more data constraining the model, give rise to improved estimation procedures. With the GP estimator, complementary observations from other cognitive tasks, beyond those collected from working memory, could be included to further constrain these models. In addition, theoretical constraints not readily encoded into linear models could also lead to better psychometric model estimates. In essence, by incorporating additional executive functions from a greater variety of testing procedures, improvements in estimation accuracy may be forthcoming, without adding more total data acquisition time for each individual.

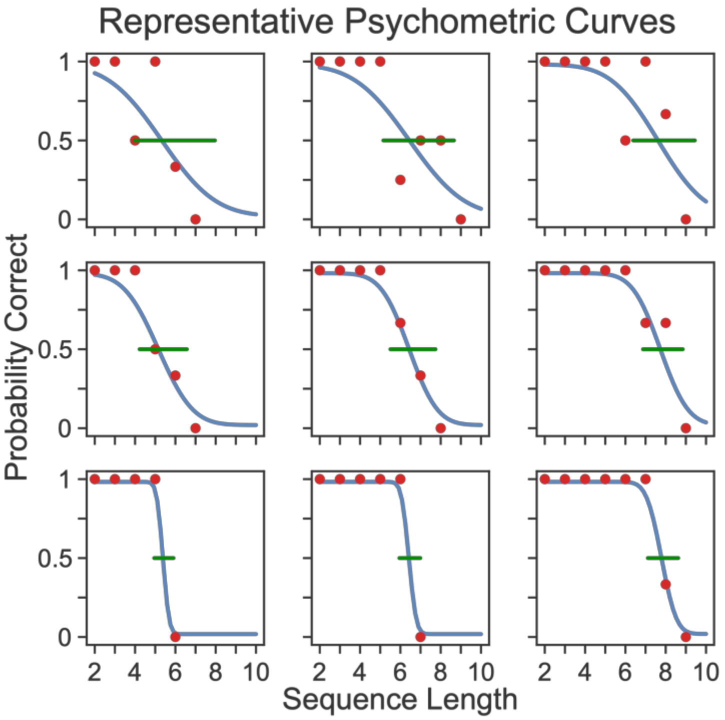

Figure 3.4: Nine representative examples of psychometric curves from the study population. Red circles indicate the proportion correct values observed at each sequence length. Blue lines indicate model fits from the data. Green lines indicate 95% credible intervals for the threshold estimates.

3.4.2 Model Validation: Constant Simulations

In the following experiments, simulations of task trial performance were carried out to evaluate the properties of the GP estimator under experimental conditions not achieved in the reference data set. The first series of simulations were designed with the constant stimuli method in order to ensure that the GP estimator delivered accurate psychometric function estimates when data quantity is not constrained. It should be noted that the advantage of using the method of constant stimuli in the form of simulations lies with its reliability, effectiveness and accuracy. In this experimental design, a fixed set of task items was repeatedly presented in random order, so as to determine the threshold of a psychometric procedure. Nine generative models, whose parameters were set fixed based on the nine representative model fits, were used to generate 100 observations at each integer-valued sequence length from 2 to 10 (i.e., 900 total observations) for each individual. This procedure ensured the inclusion, for all participants, of designated sequence lengths that they are expected to always recall successfully, sequence lengths near the threshold where predicted performance is more uncertain, and sequence lengths that participants are expected to fail at recalling almost every time.

The generative models determine the corresponding success probabilities of the simulated trials. The observations are Bernoulli distributed, ensuring that no two simulations delivered identical data for the repeated experiments. Thus, estimated psychometric fits to the data for each simulated individual were yielded by the 4-parameter GP model and compared to the ground truth psychometric curves to test the GP classifier. In Figure 3.5, one can see the resulting curves and notice that the estimated threshold and spread values were close to the ground truth values of the generative models. The value of 0.0173 for the mean RMSE between the estimates and ground truths at integer sequence lengths was considerably smaller than the smallest RMSE value for the human data. In this oversampled case, the modeled functions accurately matched the ground truth functions, which is exactly the behavior that was expected with nearly 2 orders of magnitude more data.

The constant stimuli method is a widely known and easily understood method for psychometric function estimation that was well-suited for computerized validation of our GP framework. With real human participants, sampling uniformly across all sequence lengths is neither practical, due to the time-consuming acquisition of a large quantity of trials, nor optimal, given that trials near the threshold are significantly more informative. In a real-world scenario, adaptive methods would clearly be preferred because they continually update a threshold estimate and are therefore able to select more informative samples. The purpose of testing with this simplified sampling scheme was just to provide a fixed reference against which to evaluate model performance. In summary, the method of constant stimuli was used at high sample counts to analyze the convergent performance of the GP model when data quantity grows to large amounts while ensuring fair comparisons across the different individuals.

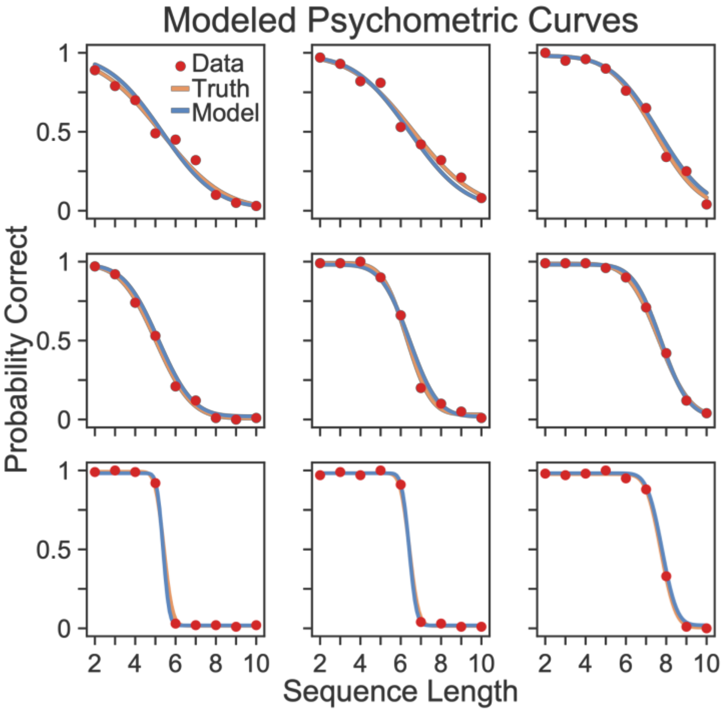

Figure 3.5: Nine simulations of 900 observations each, generated from nine associated ground truth generative models, along with associated model fits. Red circles indicate the proportion correct values observed at each sequence length. Blue lines indicate model fits from the data. Orange lines indicate ground truth.

For the 9 representative individual generative models, Figure 3.6 portrays a comparison between the threshold values resulting from the constant stimuli simulations and the ground truth values. This figure also includes representations of two other sampling schemes (adaptive staircase and actively learned observations) that will be explained in more detail in the next subsections. For constant stimuli, the mean signed threshold difference (estimate – truth) was $6.34\times10^{-4} \pm 0.16$. Because the mean estimated threshold values upon repeated simulations were close to the ground truth values, it can be argued that the threshold bias of this estimator is low. Threshold estimation variability for the full range of thresholds in the population under these sampling conditions was quantified by a mean threshold variance of

0.06 ± 0.15 for these nine examples. The novel estimator design appears to achieve accurate estimates for the psychometric thresholds present in this study population, provided that enough data is available. Since it was shown in Figure 3.5 that the novel GP model was able to capture psychometric trends accurately, based on the close correspondence between ground truth and the model curves, the satisfactory results in threshold estimates come as no surprise.

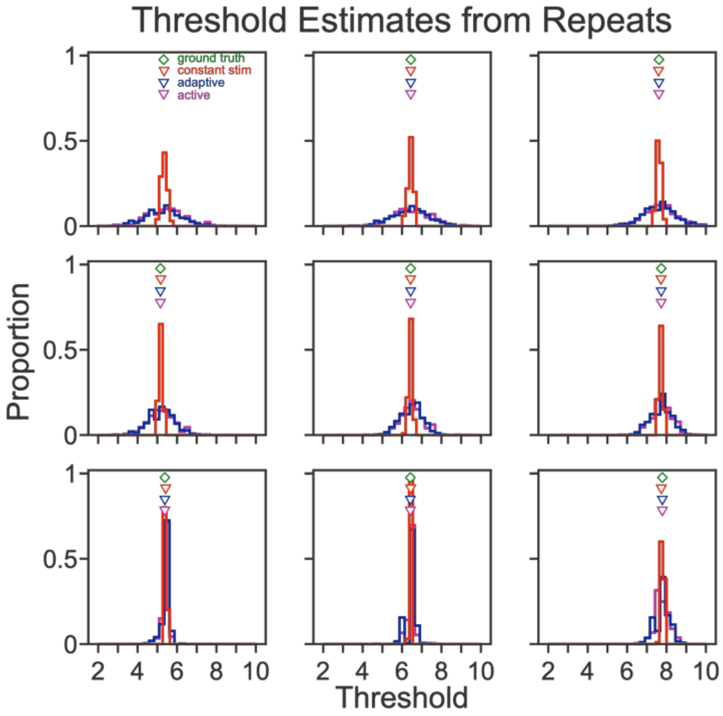

Figure 3.6: Average psychometric threshold estimation accuracy for repeated simulations. Ground truth values are indicated by green diamonds. Mean thresholds estimated by constant stimulus, adaptive staircase, and actively learned observations are indicated by red, blue and purple triangles, respectively. Overall distributions are indicated by line histograms of the same colors.

Under the same constant stimulus conditions, the threshold estimation procedure described above was repeated with several amounts of simulated data to determine the number of observations required to achieve reliable results. Even though some examples required fewer samples than others for adequate convergence, the results in Figure 3.7 showed that mean threshold estimates converged at low bias under all data conditions, with steadily diminishing variance as more data were acquired. Overall, the constant stimuli simulations were useful to show low-bias convergence of the GP estimator to the ground truth generative models when many observations are simulated at each sequence length.

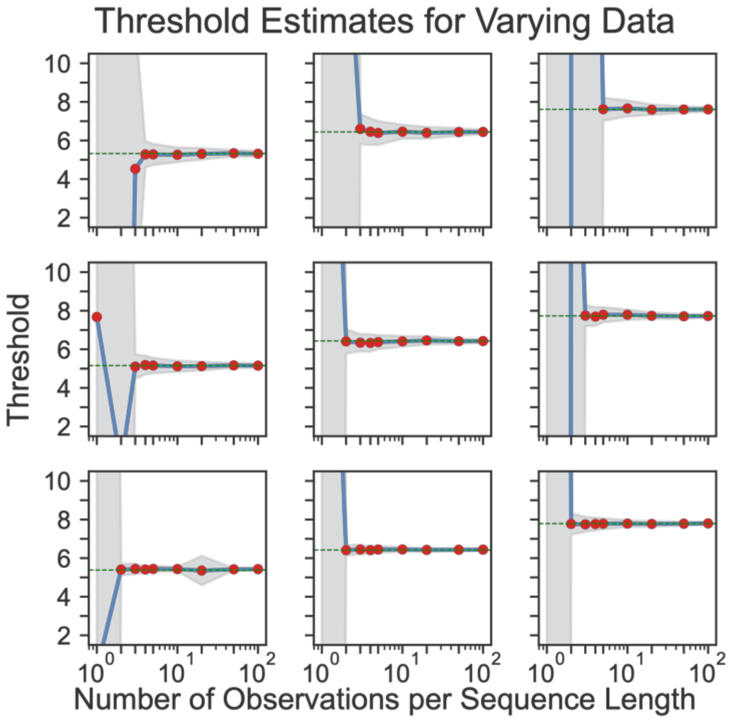

Figure 3.7: Average estimator performance as progressively more constant stimuli data are used for training (red circles, blue line) and standard deviation (shaded area).

3.4.3 Model Validation: Adaptive Staircase Simulations

The original data from the human participants was collected with an adaptive or up-down procedure, which is a variation of the method of limits (Levitt, 1971). This method is threshold-seeking, meaning that it is more efficient at estimating thresholds accurately than constant stimuli. For an equivalent trial count, adaptive acquisition places a larger fraction of trials near threshold. Since those trials are more informative for psychometric model fitting, greater accuracy is expected with this method. The second condition of simulations used a large number of repeats in order to determine estimation accuracy and reliability under the adaptive staircase procedure. Concretely, 1000 simulations were computed for each of the 9 representative models, following the same up-down procedure described above for actual human data acquisition. These simple span unidimensional detection task simulations were run with identical termination criteria to the original data collection procedures. Given their generative models, ground truths for those individuals were known beforehand, and were compared to the estimated psychometric threshold and spread from the GP model fits. The estimates from the simulated sessions were acquired with practical amounts of data, on the order of 10 to 20 adaptive sequences selected by tracking the most recent response. The results of the 9000 total simulations pointed to a reliable estimation procedure and can be visualized in Figure 3.6 above, as well as Figure 3.8 below.

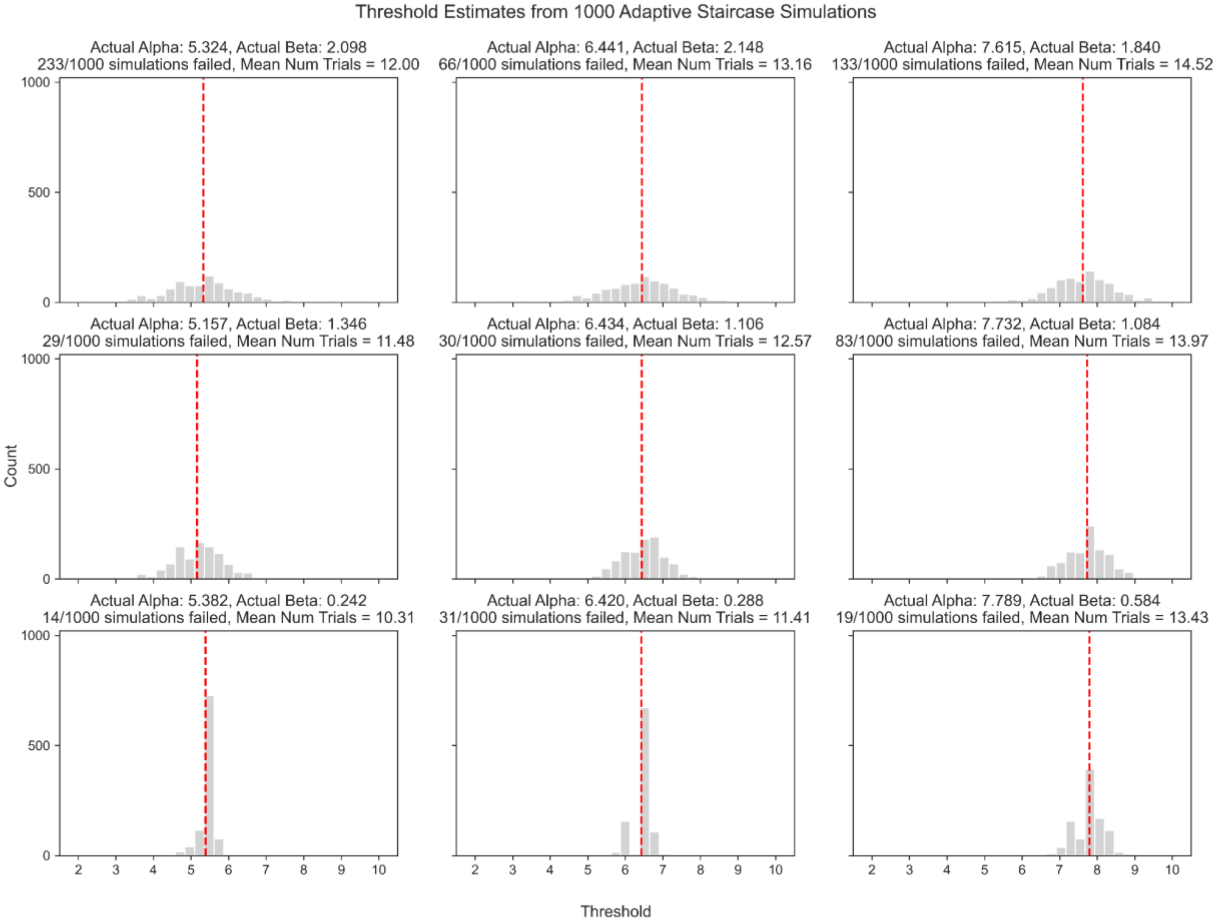

Figure 3.8: 1000 adaptive trials for each individual, terminating with the same criteria used for collection of real data from the study population.

Average trial count was 12.5 ± 2.04 (mean ± standard deviation) for the simulation results, which compares with the average trial count for the human data results: 12.4 ± 1.87 for the entire study population, and 12.8 ± 1.99 for the 9 representative individuals. The mean signed threshold difference (estimate – truth) was 3.41×10–3 ± 0.59, with a threshold difference variance of 0.41 ± 0.32. Variance was relatively small (similar to 30–50 constant stimulus trials) and the threshold values are low, too, indicating that threshold bias is small under these reduced data conditions. Therefore, the GP estimator can be used to fit psychometric functions reasonably well by altering the next sample based on the result of the previous sample. Nonetheless, adaptive data acquisition is still not optimal and can be further constrained to estimate psychometric functions even more efficiently with the GP model.

3.4.4 Model Validation: Active Learning Simulations

Active data acquisition is a particular form of adaptive acquisition in which subsequent test items are selected to optimize an objective function incorporating all previous data. Model updates after each new observation ensure that this optimization stays current as new data are collected. To evaluate performance of the GP estimator when model information gain is maximized, active learning simulations were run.

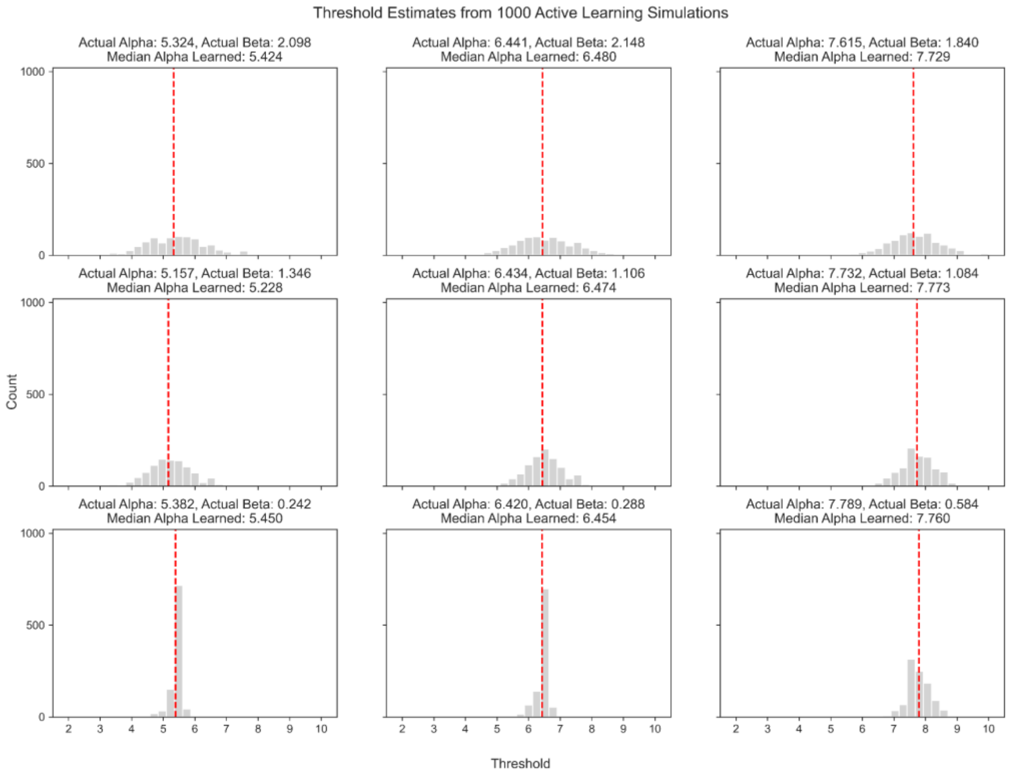

Figure 3.9: 1000 active trials for each individual, terminating after the average number of trials obtained using the simulated adaptive trials.

Figure 3.9 shows the threshold estimation results of active data acquisition for the 9 representative models given those simulated individuals' previous trials from the working memory task. These simulations were terminated after the same mean number of adaptive trials for each individual, rounded up. Mean trial counts across all 9 models under these conditions were 13.4 ± 1.50. With respect to adaptive or constant stimulus procedures, the small variation of the actively sampled model parameters demonstrated greater accuracy and reliability of the active learning design.

3.5 Discussion

This chapter introduced a non-parametric Gaussian Process probabilistic classifier that was capable of performing inference with fully predictive models that estimate the entire psychometric function. This promising GP implementation for 4-parameter psychometric functions is a crucial foundation to add more task variables and incorporate additional informative predictors. As a Bayesian method, it has the ability to couple any existing theoretical modeling constraints with any empirical data constraints that might be available. The purpose of expanding the GP kernel into new dimensions and increasing model complexity is to accomplish higher dimensional estimates of psychometric functions. Multiple experiments were presented above to demonstrate the results of the novel development in a working memory capacity task that was used as a test case. The goal of this research line, however, is to construct, apply, and generalize similar scalable models to other applications in psychometrics, too. The validity of the spatial working memory task is anticipated to extend to a variety of cognitive and perceptual domains.

In previous GP frameworks, a Bernoulli likelihood function was the only likelihood available and repeated test items had not been accommodated. Thereby, a key expansion of this work was the novel binomial likelihood function that was developed and validated. It is worth underlining that the likelihood function was driven by a generalized 4-parameter sigmoid, instead of a 2-parameter cumulative Gaussian sigmoid. Performance was modeled as a function of the independent variable "sequence length", equivalent to "task difficulty", so active learning was feasible for this span task. By testing at the locations that corresponded to higher uncertainty in the posterior distribution, active learning was implemented for the new binomial likelihood, whereby prospective data can be collected optimally. In other words, a more advanced data collection shortcut than adaptive staircases was shown to be successfully implemented to actively acquire the data. The unidimensional generalization to the 4-parameter model opens up a much larger variety of problems that could benefit from the efficiency gains of actively learned GP estimation. Future experimental designs for working memory assessment could benefit from a real-valued input variable to provide higher resolution options for active learning to select among. This could be achieved by replacing "sequence length", which only takes integer values, with a designed variable of "task difficulty" for real-valued scores.

For the working memory test case, the initial focus was placed on accuracy scores from young adults as well simulation results. The purpose of these evaluations was to validate the human dataset available and verify that the overall accuracy scores were consistent with previous results. The cognitive models were tested under constant stimuli, adaptive staircase and active learning scenarios. In these settings, the efficiency gains were not extraordinary with respect to generalized logistic regression models. This finding is consistent with the notion that the power of the GP framework is its scalable nature, yielding substantial efficiency gains upon the addition of more latent variables with complex interactions (Song et al., 2018). The successful outcome of the current work in a unidimensional space was that the GP was capable of performing high-quality fits, reflecting the anticipated psychometric threshold and spread trends. Estimator bias and variance (i.e., reliability) were both low as long as sufficient data was acquired, which fell in the range of 30-50 total data points for threshold and 100-200 total data points for spread. Threshold estimates appeared to be less variable under repeat adaptive staircase testing conditions (i.e., more reliable) than the traditional metric of maximum sequence length, which appeared to be biased toward higher values. Overall, however, the two metrics were shown to be highly correlated, implying that thresholds estimated using the GP method would be at least as informative as maximum sequence length recalled.

In essence, the focus of the current study has been to verify that the proposed modeling procedure accurately captures trends in the data. A working memory task was selected for the purposes of testing and validation. Using modern machine learning algorithms, the GP framework yielded probabilistic cognitive models that represented psychometric performance data accurately and consistently. The GP models have been designed to measure executive functions with limited reference to other individuals' performance, promoting more equitable behavioral models. This GP based modeling procedure has the potential to be widely applicable to quantify psychological variables in a vast array of cognitive and perceptual tasks. Future analyses could involve demonstrating its functionality in other cognitive domains outside of memory.

Chapter 4: Conclusion

The target of the current research is to determine on a regular basis if the current cognitive performance of an individual is significantly altered from a baseline level. Despite the enormous advances in our understanding of executive functions, there is still plenty of scope for improving upon conventional testing methods designed to draw inferences about them. Novel techniques must be developed to conduct more nearly optimal evaluations of intrasubject variability or underlying latent variables over time with increased inferential power. In general, executive functioning testing requires querying participants multiple times to accomplish model fits to the behavioral data and ultimately better understand the corresponding underlying latent constructs. Behavioral assessments that acquire data actively provide enhanced efficiency and thus avoid large numbers of tedious trials, both in unidimensional and multidimensional feature spaces. Active testing modalities are especially compelling for estimators that build model fits with limited reference to other individual's performance. In view of these facts, this thesis proposes to develop scalable testing strategies that are equitable, less data-intensive and that take advantage of test-specific active sample collection.

Undoubtedly, the burdensome time requirement of data collection is a notable shortcoming of current testing procedures. One should note that it can be more challenging to shorten timing and accuracy tasks that do not have any independent variables. The main issue with non-threshold-seeking behavioral tests where the overall difficulty of the set of trials is the same for all individuals is that one cannot select the most informative elements within the independent variables, as they are simply non-existent. Fortunately, researchers are given the opportunity to resort to sequential testing for improved efficiency and a substantial speed up of these types of cognitive tasks. Chapter 2 has shown that a notable reduction in the number of trials presented to participants can be accomplished with sequential tests that can reliably detect the presence of significant differences in latent constructs. Perhaps surprisingly, most current designs do not take advantage of tools like sequential testing, despite its potential benefits. Hence, this thesis aimed to act upon the lack of flexibility of current unidimensional tests by attempting to demonstrate the feasibility and efficiency of sequential testing schemes to scale down the data requirements needed to estimate fluctuations in executive function processes. The empirical experiments described above for tasks without independent variables illustrated that more efficient determinations about fluctuations in executive functions can be made while controlling for error rates and saving valuable resources, such as time and money, when opting to adopt sequential testing. The general principles of sequential testing can be applied to other cognitive and perceptual domains, so researchers are encouraged to consider putting this strategy forward in order to speed up evaluations in their particular scientific inquiries.

Multiple routes exist to extend the current non-threshold-seeking sequential testing framework. The first extension to point out is that timing models could include not only reaction time data, but also accuracy information. Although keeping track of changes in accuracy of the responses might not be as impactful in tasks like Numerical Stroop or Countermanding, it could still be a valuable feature to detect differences in executive functions with even fewer observations. Similarly, models for accuracy tasks that usually do not take the amount of time employed to respond into account could be expanded to include this additional feature. All in all, quicker determinations about cognitive states could potentially be accomplished in non-threshold-seeking tasks by testing for fluctuations in reaction time and accuracy simultaneously, i.e., extending to a multidimensional model using data streams from just the test in question.

Furthermore, while this study demonstrated the benefit of a group sequential approach for cognition in a particular experimental setting, it is possible that sequential designs could reflect even more significant resource savings in other behavioral tasks or under slightly different simulation conditions. Future analyses might consider design modifications that include the use of a variety of alpha-level corrections (for instance, O'Brien–Fleming or Pocock corrections), incorporation of variable futility stopping criteria (like a beta-spending function), as well as combinations of several numbers of interim steps and batch sizes. A logical next step beyond frequentist strategies comprises the replication of the sequential approach using Bayesian statistical tools, which are able to put to use information from earlier stages of the study. That is, making use of Bayesian statistics, researchers could take advantage of pre-study prior information about the measures of interest to perform model comparison and derive credible intervals for estimates.

Chapter 3 turns the attention to threshold-seeking assessments, proving that machine learning tools can greatly assist in building more advanced models of psychometric tasks that are designed with the inherent multidimensional nature of real tasks in mind. The result of this chapter is the introduction of the first estimator that is capable of successfully approximating full psychometric functions (i.e., all 4 parameters) with a novel binomial likelihood that accommodates testing at repeated discrete values. The modern GP framework presented by this research line not only worked as designed in a unidimensional simple span test case, but also has a large potential for extension that is simply nonexistent in conventional testing frameworks. The current study found clear support for the successful quantification of psychometric task performance using a GP based modeling procedure that accurately captures trends in the data without referencing individuals against one another and norming the results. In addition, this method provided interval estimates of model parameters that are useful to perform hypothesis testing and report the degree of confidence to which working memory in an individual at a particular time point is significantly different from baseline. The active testing capability was encoded in the GP model by actively querying at the sequence of input values that minimizes the model's posterior uncertainty about its prediction. Broadly, equivalent active learning strategies could be implemented for threshold-seeking tasks that have independent variables like task difficulty.

To conclude, batteries of unidimensional tests are typically used in behavioral science to quantify the executive functions of working memory, cognitive flexibility, and inhibitory control, but correlations within tasks in a test battery are not exploited. This extension remains a subject for future studies. The logical path of cognitive and perceptual testing methods is expected to head towards the creation of models that can capture interrelationships among each test, which may be nonlinear. It is important to recognize the need to extend the current work in order to build estimators that profit from the correlations among the individual model outputs for each test. As a final remark, it is vital to mention that latent variable models are anticipated to be an incredibly useful tool to conjointly estimate multiple variables. Ideally, latent variable models could uncover the relationships between latent task abilities, utilizing data originating from one test to improve model estimates for another test. Scalability is a very relevant feature that is predominantly absent in current methods and that this study contributes to incorporate. Therefore, subsequent investigations are needed to work towards expanding the dimensionality of the underlying Gaussian process to incorporate multiple tests and prior beliefs using previous individuals who have received combinations of the tests in question.

References

Barbour, D. L. (2019). Precision medicine and the cursed dimensions. *Npj Digital Medicine*, *2*(1), 4. https://doi.org/10.1038/s41746-019-0081-5

Barbour, D. L., DiLorenzo, J. C., Sukesan, K. A., Song, X. D., Chen, J. Y., Degen, E. A., Heisey, K. L., & Garnett, R. (2019). Conjoint psychometric field estimation for bilateral audiometry. *Behavior Research Methods*, *51*(3), 1271–1285. https://doi.org/10.3758/s13428-018-1062-3

Brickenkamp, R., & Zillmer, E. (1998). *The d2 test of attention*. Hogrefe & Huber.

Brucki, S. M. D., & Nitrini, R. (2008). Cancellation task in very low educated people. *Archives of Clinical Neuropsychology*, *23*(2), 139–147. https://doi.org/10.1016/j.acn.2007.11.003

Brunetti, R., Del Gatto, C., & Delogu, F. (2014). eCorsi: Implementation and testing of the Corsi block-tapping task for digital tablets. *Frontiers in Psychology*, *5*, 939. https://doi.org/10.3389/fpsyg.2014.00939

Buss, E., Hall III, J. W., & Grose, J. H. (2006). Development and the role of internal noise in detection and discrimination thresholds with narrow band stimuli. *J Acoust Soc Am*, *120*(5), 2777–2788.

Chen, J. (2020). A Generalized Gaussian Process Likelihood for Psychometric Function Estimation. *Engineering and Applied Science Theses & Dissertations*. https://doi.org/10.7936/tabn-yr37

Cheng, C.-A., & Boots, B. (2017). Variational inference for Gaussian process models with linear complexity. *ArXiv Preprint ArXiv:1711.10127*.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. https://doi.org/10.3758/bf03196772

Corsi, P. M. (1972). *Human memory and the medial temporal region of the brain.* [McGill University]. https://escholarship.mcgill.ca/downloads/4m90dw30g

Cortés Pascual, A., Moyano Muñoz, N., & Quílez Robres, A. (2019). The Relationship Between Executive Functions and Academic Performance in Primary Education: Review and Meta-Analysis. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/article/10.3389/fpsyg.2019.01582

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*(11), 2037–2078. https://doi.org/10.1016/j.neuropsychologia.2006.02.006

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Dodge, H. F., & Romig, H. G. (1929). A method of sampling inspection. *The Bell System Technical Journal*, *8*(4), 613–631. https://doi.org/10.1002/j.1538-7305.1929.tb01240.x

Eggen, T. J. H. M. (1999). Item Selection in Adaptive Testing with the Sequential Probability Ratio Test. *Applied Psychological Measurement*, *23*(3), 249–261. https://doi.org/10.1177/01466219922031365

Etz, A. (2018). Introduction to the Concept of Likelihood and Its Applications. *Advances in Methods and Practices in Psychological Science*, *1*(1), 60–69. https://doi.org/10.1177/2515245917744314

Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., & Fair, D. A. (2019). The Heterogeneity Problem: Approaches to Identify Psychiatric Subtypes. *Trends in Cognitive Sciences*, *23*(7), 584–601. https://doi.org/10.1016/j.tics.2019.03.009

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

García-Pérez, M. A., & Alcalá-Quintana, R. (2005). Sampling plans for fitting the psychometric function. *The Spanish Journal of Psychology*, *8*(2), 256–289. https://doi.org/10.1017/s113874160000514x

Gardner, J. M., Malkomes, G., Garnett, R., Weinberger, K. Q., Barbour, D., & Cunningham, J. P. (2015). Bayesian active model selection with an application to automated audiometry. *Adv Neural Inf Process Syst*, 2377–2385.

Giacalone, M., Agata, Z., Cozzucoli, P. C., & Alibrandi, A. (2018). Bonferroni-Holm and permutation tests to compare health data: Methodological and applicative issues. *BMC Medical Research Methodology*, *18*(1), 81. https://doi.org/10.1186/s12874-018-0540-8

Gold, J. I., & Ding, L. (2013). How mechanisms of perceptual decision-making affect the psychometric function. *Progress in Neurobiology*, *103*, 98–114. https://doi.org/10.1016/j.pneurobio.2012.05.008

Gronwall, D. M. A., & Sampson, H. (1974). *The psychological effects of concussion*. Auckland University Press ; Oxford University Press.

Hart, A. (2001). Mann-Whitney test is not just a test of medians: Differences in spread can be important. *BMJ : British Medical Journal*, *323*(7309), 391–393.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and

Consequences of P-Hacking in Science. *PLOS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Hughes, C., & Graham, A. (2002). Measuring Executive Functions in Childhood: Problems and Solutions? *Child and Adolescent Mental Health*, *7*(3), 131–142. https://doi.org/10.1111/1475-3588.00024

Kingdom, F. A. A., & Prins, N. (2010). *Psychophysics: A Practical Introduction*. Elsevier.

Kleinert, C., Christoph, B., & Ruland, M. (2021). Experimental Evidence on Immediate and Long-term Consequences of Test-induced Respondent Burden for Panel Attrition. *Sociological Methods & Research*, *50*(4), 1552–1583. https://doi.org/10.1177/0049124119826145

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *J Vis*, *5*(5REVIEWED), 8.

Lakens, D., Pahlke, F., & Wassmer, G. (2021). *Group Sequential Designs: A Tutorial*. PsyArXiv. https://doi.org/10.31234/osf.io/x4azm

Larsen, T., Malkomes, G., & Barbour, D. (2021). Accelerating Psychometric Screening Tests with Prior Information. In A. Shaban-Nejad, M. Michalowski, & D. L. Buckeridge (Eds.), *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability* (pp. 305–311). Springer International Publishing. https://doi.org/10.1007/978-3-030-53352-6_29

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2), Suppl 2:467+.

Martínez, F., Ramos-Ortega, M., & Vila, J. Ó. (2018). Executive efficacy on Stroop type interference tasks. A validation study of a numerical and manual version (CANUM). *Anales de Psicologia*, *34*, 184–196. https://doi.org/10.6018/analesps.34.1.263431

Matthews, A. G. de G., Hensman, J., Turner, R., & Ghahramani, Z. (2016). On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. *Artificial Intelligence and Statistics*, 231–239.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Morein-Zamir, S., Nagelkerke, P., Chua, R., Franks, I. M., & Kingstone, A. (2004). Inhibiting prepared and ongoing responses: Is there more than one kind of stopping? *Psychonomic Bulletin and Review*, *11*(6), 1034–1040. https://doi.org/10.3758/BF03196733

Neumann, K., Grittner, U., Piper, S. K., Rex, A., Florez-Vargas, O., Karystianis, G., Schneider, A., Wellwood, I., Siegerink, B., Ioannidis, J. P. A., Kimmelman, J., & Dirnagl, U. (2017). Increasing efficiency of preclinical research by group sequential designs. *PLOS Biology*, *15*(3), e2001307. https://doi.org/10.1371/journal.pbio.2001307

Pahor, A., Mester, R. E., Carrillo, A. A., Ghil, E., Reimer, J. F., Jaeggi, S. M., & Seitz, A. R. (2022). UCancellation: A new mobile measure of selective attention and concentration. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01765-5

Park, J., & Haran, M. (2018). Bayesian Inference in the Presence of Intractable Normalizing Functions. *ArXiv:1701.06619 [Stat]*. http://arxiv.org/abs/1701.06619

Quiñonero-Candela, J., & Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *J Mach Learn Res*, *6*(DecEDITED), 1939–1959.

Ramani, G. B., Daubert, E. N., Lin, G. C., Kamarsu, S., Wodzinski, A., & Jaeggi, S. M. (2020). Racing dragons and remembering aliens: Benefits of playing number and working memory games on kindergartners' numerical knowledge. *Developmental Science*, *23*(4), e12908. https://doi.org/10.1111/desc.12908

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Reynolds, C. R., & Suzuki, L. A. (2012). Bias in Psychological Assessment. In *Handbook of Psychology, Second Edition*. American Cancer Society. https://doi.org/10.1002/9781118133880.hop210004

Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/article/10.3389/fpsyg.2019.00813

Shen, Y. (2013). Comparing adaptive procedures for estimating the psychometric function for an auditory gap detection task. *Atten Percept Psychophys*, *75*(4), 771–780.

Song, X. D., Sukesan, K. A., & Barbour, D. L. (2018). Bayesian active probabilistic classification for psychometric field estimation. *Attention, Perception & Psychophysics*, *80*(3), 798–812. https://doi.org/10.3758/s13414-017-1460-0

Spiegelhalter, D. (2003). Risk-adjusted sequential probability ratio tests: Applications to Bristol, Shipman and adult cardiac surgery. *International Journal for Quality in Health Care*, *15*(1), 7–13. https://doi.org/10.1093/intqhc/15.1.7

Strasburger, H. (2001). Converting between measures of slope of the psychometric function. *Percept Psychophys*, *63*(8EDITED), 1348–1355.

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—Or Why the P Value Is Not Enough.

Journal of Graduate Medical Education, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Tombaugh, T. N. (2006). A comprehensive review of the Paced Auditory Serial Addition Test (PASAT). *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *21*(1), 53–76. https://doi.org/10.1016/j.acn.2005.07.006

Towse, J. N., Hitch, G. J., & Hutton, U. (1998). A reevaluation of working memory capacity in children. *Journal of Memory and Language*, *39*(2), 195–217.

Treutwein, B., & Strasburger, H. (1999). Fitting the psychometric function. *Percept Psychophys*, *61*(1), 87–106.

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. (2014). A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research. *Child Development*, *85*(3), 842–860. https://doi.org/10.1111/cdev.12169

Viana-Sáenz, L., Sastre-Riba, S., Urraca-Martínez, M. L., & Botella, J. (2020). Measurement of Executive Functioning and High Intellectual Ability in Childhood: A Comparative Meta-Analysis. *Sustainability*, *12*(11), 4796. https://doi.org/10.3390/su12114796

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313. https://doi.org/10.3758/BF03194544

Wiens, A. N., Fuller, K. H., & Crossen, J. R. (1997). Paced Auditory Serial Addition Test: Adult norms and moderator variables. *Journal of Clinical and Experimental Neuropsychology*, *19*(4), 473–483. https://doi.org/10.1080/01688639708403737

Williams, C. K. I. (1998). Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M. I. Jordan (Ed.), *Learning in Graphical Models* (pp. 599–621).

Springer Netherlands. http://dx.doi.org/10.1007/978-94-011-5014-9_23

Yssaad-Fesselier, R., & Knoblauch, K. (2006). Modeling psychometric functions in R. *Behavior Research Methods*, *38*(1), 28–41. https://doi.org/10.3758/bf03192747

Zelazo, P. D., Blair, C. B., & Willoughby, M. T. (2016). Executive Function: Implications for Education. NCER 2017-2000. In *National Center for Education Research*. National Center for Education Research. https://eric.ed.gov/?id=ED570880

Żychaluk, K., & Foster, D. H. (2009). Model-free estimation of the psychometric function. *Attention, Perception & Psychophysics*, *71*(6), 1414–1425. https://doi.org/10.3758/APP.71.6.1414