Applying a hierarchical linear model to investigate intra-individual variability in response time as a predictor for neurocognitive impairment

Bruno Hidalgo Monroy Lerma
The Wheeler Lab
05/02/2022

Approved by

Mark E. Wheeler, Ph.D. (Faculty Mentor)   Michael D. Hunter, Ph.D. (Second Reader)

Table of Contents

0. Specific Aims

Efforts to treat Alzheimer's disease and other forms of dementia continue to make exciting progress. However, the absence of approved effective treatment options available to the exponentially growing number of patients is an increasingly deafening silence.

This study contributes to the alternative strategy of developing sensitive diagnostics that allow clinicians to prevent the onset of severe symptoms altogether. It does so by applying a comprehensive series of neurocognitive assessments to old age and young age cohorts to produce profiles of neurocognitive health for each individual. Individuals also undergo a memory recall task where response time variability is recorded. Data from the neurocognitive assessments will be incorporated into a hierarchical linear model to investigate their relation to intra-individual variability in response time whilst accounting for their nested data structure. In this way, I contribute to the investigation of intra-individual variability in response time as a potential cognitive marker of early-stage cognitive impairment.

1. Introduction

1.1 Dementia: a growing problem

In 2013, when there were an estimated 5.0 million individuals aged 65 years or older diagnosed with dementia or Alzheimer's disease (AD) in the United States, Hebert et al. (2013) predicted that this figure would grow to 5.8 million by 2021. Alarmingly, they also reported a projected tripling of this figure by 2050. In fact, their model has underestimated the growth in cases thus far – with the current American population living with dementia or AD estimated at 6.2 million (n.a., 2021).

The mission to treat and cure dementia has seen extensive funding and inspiring innovation, yet there are few successful treatments to show for these efforts (Yiannopoulou, 2020). Many have argued that this shortcoming is not attributed to an incorrect approach to the problem, but rather an approach to the incorrect problem (Opar, 2010). Current research endeavors attempt to treat neurocognitive damage at an advanced stage through disease-modifying techniques, leading to daunting tasks such as the reversal of protein accumulation or neural degeneration (Yiannopoulou, 2020). However, the task would become much easier if we could avoid these and other disease mechanisms altogether. Converging evidence demonstrates that the pathological substrates of dementia and AD begin to develop decades before overt symptoms arise (Opar, 2010). Accordingly, the Alzheimer's Association Research Roundtable has urged research to address the alternative problem of developing more sensitive diagnostic methods (Christ et al., 2018). Such methods promise to identify neurocognitive disorders in earlier stages and empower clinicians to develop treatment plans that prevent – or meaningfully delay – the onset of overt symptoms.

A diagnostic-based response to the projected increase in dementia cases will only be effective through mass screening, and corresponding logistical barriers ought to be considered from the beginning of the research process. Diagnostic methods dependent on expensive techniques or inaccessible materials run the risk of not only failing to screen a sufficient portion of the population, but also exacerbate socioeconomic disparities by systemically excluding certain demographics from screening. Because of their behavior- and task-based acquisition, cognitive signatures hold potential for low-cost, mass-reproducible diagnostic techniques. Several papers have reported that cognitive deficits can be detected several years before dementia is clinically diagnosed, and thus can be used for the early detection or prediction of dementia (Linn et al., 1995; Locascio et al., 1995; Pasquier, 1998; Alves et al., 2018).

1.2 Intra-individual variability

Intra-individual variability (IIV) in repeated cognitive measures has received growing attention in the field of cognitive neuroscience because of its frequently reported relation with neurocognitive impairment (MacDonald et al., 2008). Latency-based cognitive measures – where the time taken to respond is a relevant factor – are particularly useful for the investigation of IIV, as they yield a larger range than other cognitive scores. Because of its latency-based nature, response time has been widely used in the literature and noted to be particularly sensitive to individual performance differences (Christ et al., 2018). However, most studies investigating IIV in response time (IIV-RT) have not accommodated the data's structure in the most effective way. Experimental paradigms investigating IIV-RT often yield an inherently nested – or hierarchical – data structure, where the cognitive measures produced by an individual are nested within the individual (that is, they belong to the individual). Researchers have classically dealt with this nested data structure by aggregating the data (i.e., reducing the repeated measures to an average data point that represents the individual), thereby losing rich information about the details of an individual's performance (Woltman et al., 2012).

This study will elucidate the neurocognitive processes that determine IIV-RT through a comprehensive neurocognitive battery, and to incorporate this data into a hierarchical linear model that accurately accounts for nesting. In doing so, we will investigate the relation between IIV-RT, neurocognitive health, and age more effectively – taking a step towards the development of more sensitive and accessible diagnostic measures for the detection of early-stage cognitive impairment.

2. Literature Review

2.1 Cognitive impairment & intra-individual variability

As converging evidence continues to support the notion that the biological mechanisms of AD and other forms of dementia begin to occur several decades before the onset of overt cognitive and functional changes, the importance of early intervention becomes increasingly evident. Correspondingly, the Alzheimer's Association Research Roundtable has stressed the urgent need for more sensitive diagnostic methods that can detect and perhaps predict cognitive and functional changes at the earliest stages of the disease mechanism (Snyder et al., 2014). Such tools will empower clinicians to prevent or delay neurocognitive damage and disease acceleration, rather than attempting to reverse complex disease pathways or merely manage symptoms. Converging evidence demonstrates that cognitive deficits can arise years before the clinical diagnosis of dementia, implying their usefulness for earlier detection (Linn et al., 1995; Locascio et al., 1995; Pasquier, 1998; Alves et al., 2018). Correspondingly, the development of diagnostic techniques that do not necessitate expensive materials and devices is feasible.

Over the past two decades, IIV in cognitive measures has gained traction as a measure with potential for diagnostic application due to its relation with cognitive impairment in both healthy and dementia-diagnosed individuals. MacDonald et al. (2008) define IIV as "lawful but transient within-person changes in performance, such as trial-by-trial fluctuations on a reaction time task". The investigation of IIV represents a shift in the literature for dementia research from the classical approach of examining group-level differences to the analysis of individual-level trends. Li et al. (2017) stressed the importance of individual-level analysis, given the marked heterogeneity of cognitive functioning in late adulthood and old age.

This sentiment resonates with other researchers in the field, as IIV has been investigated through the variability of a variety of measures including trial-by-trial accuracy and day-by-day cognitive test scores. However, Christ et al. (2018) argued that "latency-based measures such as reaction time are particularly well-suited to IIV research because they have larger ranges than traditional cognitive test scores, thus making them more sensitive than traditional cognitive tests to individual performance differences." Additionally, reaction time-based tasks are easily scalable to gather many samples of performance and, in this way, are less sensitive to re-test effects (Christ et al., 2018). As such, much of the literature regarding IIV employs IIV in reaction (or response) time (IIV-RT), and this will be the primary measure of interest in the present study. The remainder of this literature review will thus focus on IIV-RT.

It should be noted that IIV can be categorized into four main categories. Mella et al. (2013) describe these as, "Short-term or trial-to-trial, within-task variability, or *inconsistency*, denoting transient and rapid fluctuations that occur over short-term time scales… Intra-individual variability across tasks, or *dispersion*… Relatively permanent alterations that evolve slowly over relatively long-term time scales, through training, or development, that is, *intraindividual change*… [and] Interindividual or between-individual variability, also termed *diversity.*" It is critical to keep the distinction between these categories in mind when reading relevant literature, as reports relating to one category of IIV may not necessarily translate to another. IIV-RT is best described as 'inconsistency', and Mella et al. (2013) reinforced the repeatedly reported finding that inconsistency in performance is associated with both functional and structural neural substrates of cognition that are more statistically significant than the other categories of IIV. Given the noted greater statistical significance of its relation with relevant variables, the present study will explicitly investigate inconsistency in response time.

2.2 Structural & functional determinants of IIV-RT

Mella et al. (2013) reported a noteworthy relation between IIV-RT and white matter integrity through diffusion tensor imaging (DTI) – an established method for the imaging of white matter. This study is particularly comprehensive because it also integrates fMRI analysis to reinforce its conclusions. The relevance of white matter integrity to cognitive functioning is echoed throughout the literature. Li et al. (2017) applied functional connectivity MRI (fcMRI) – a form of resting state fMRI – to establish whole-brain connectivity maps and relate these to cognitive ability on an individual basis. Their findings stress the crucial function of long-range and inter-network white matter projections on individual cognition, relative to local and intra-network circuits. Given the complex coordination between brain regions required for cognitive function, the relevance of white matter tracts is expected. Indeed, evidence shows that deficient white matter myelination and neural noise can disrupt the efficiency of information propagation along axons, and in this way may underlie inconsistency in neurocognitive performance (Walhovd & Fjell, 2007). Moreover, the highest degree of between-subject variability is observed in white matter projections that are relevant to cognition (Li et al., 2017). Mueller et al. (2013) similarly cited evidence of higher inter-subject variability in long association white matter tracts implicated in higher-order association and integration tasks. This variability must be accounted for during white matter analysis to avoid mistakenly describing inherent variability in healthy subjects as pathological changes. These findings also suggest the possibility of a relation between the inter-individual variability in neural structures and the intra-individual variability in neurocognitive task performance that the structures are implicated in.

2.3 Neuromodulatory determinants of IIV-RT

The relation between IIV-RT and neurocognitive health has also been approached from the molecular perspective. In their 2009 paper, MacDonald et al. described the widely studied structural and functional determinants of IIV-RT, which have been described in this literature review. However, they propose a third avenue for the investigation of IIV-RT – neuromodulation. They cite literature that demonstrated that dysfunctional systems of dopamine (DA) and acetylcholine modulation led to increased neural noise (Backman et al., 2006). MacDonald et al. (2009) specifically focused on DA, using previous papers based on neurocomputational models to support their arguments (Li & Lindenberger, 1999; Li et al., 2001). According to these models, DA facilitates responsivity of neural networks in activity transmission both between and within networks. It enhances the neural signal relative to background noise, improving the signal-to-noise ratio and promoting firing frequency in innervated neurons. Reduced DA activity results in less distinct cortical representations, which translates to impaired performance on neurocognitive assessments and increased IIV-RT.

MacDonald et al. (2009) further supported the connection between DA activity and IIV using genetic evidence related to COMT, an enzyme that degrades extracellular DA in the frontal cortex. A previous paper by Stefanis et al. (2005) showed that Val carriers, which have more active COMT and thus lower prefrontal DA activity, had higher IIV than Met carriers, who have less active COMT and greater prefrontal DA activity. This study established links between DA levels in the orbitofrontal cortex, anterior cingulate cortex, and hippocampal complex. The orbitofrontal cortex is known to be critical for decision making, while the anterior cingulate cortex is implicated in error detection, attention, conflict monitoring, and motivation. It is also noteworthy that the anterior cingulate cortex has connections to prefrontal and parietal cortices. In general, these findings offer a useful perspective in providing more comprehensive evidence for the implication of neural systems in IIV, rather than environmental or contextual factors. They also offer potentially causal relationships between neurocognitive performance and dysregulated neuromodulation. While the present study will not be able to directly explore neuromodulatory systems, this literature contributes comprehensive support for the relevance of IIV-RT in dementia research.

2.4 Contribution to the literature

Whilst the relation between IIV-RT and cognitive impairment has been substantially investigated from a variety of perspectives, there remain shortcomings that ought to be addressed moving forward. A fundamental issue with many of the papers referenced here is the way in which the data is modeled. When collecting latency-based data via a memory paradigm, a nested data structure is typically created (Bauer et al., 2013). This is not an issue unique to this field of study. The mishandling of nested data structures has arisen in various other fields where such structures are present. In these cases, this problem has been successfully addressed by applying hierarchical linear models to accommodate nesting more accurately from a statistical perspective (Hox, 2010; Woltman et al., 2012).

The present study will investigate the relation between IIV-RT and neurocognitive performance across young age (YA) and old age (OA) cohorts. Participants undergo encoding of word-picture pairs, and later engage in an associative memory retrieval task where their response latency (i.e., response time) is measured.

Though this has been previously studied, I plan to handle this dataset by applying a hierarchical linear model (HLM) in order to accurately preserve the data's nested structure at the statistical level. Specifically, the repeated measures of response time will be treated as a time variable, introducing a longitudinal modeling perspective that will maximize the number of observations of response times and more accurately capture their sequential progression (Snijders & Bosker, 2004; Hox, 2010). Through this approach, I will elucidate more sensitively the within- and between-level relations between IIV-RT, neurocognitive performance, and age.

3. Methods

3.1 Participant recruitment

Participants were recruited from a community convenience sample using informational flyers. A preliminary screening was performed to exclude participants with neurological exclusions, not within the study's age categories (21-35 and 60-75), or physical exclusions relating to data collection within a mock fMRI scanner. Neurological exclusions include non-proficiency in English, learning disabilities, language disorders, neurologic or mental illness, and history of strokes or heart attacks. Physical exclusions include claustrophobia, implanted metal devices, and weight over 250 pounds. Given that this study aims to identify indicators of cognitive impairment that precede current diagnostic tests, the data set included only cognitively healthy individuals, as determined by standard examinations.

3.2 Participant analysis & neurocognitive assessments

Participants' demographic information was recorded, including years of education (YoE), ethnicity, race, and gender. A Mini-Mental State

Examination (MMSE) and Clock Drawing Test was performed as further screening for cognitive impairment that would exclude the participant from the study (Tombaugh & McIntyre, 1992; Aprahamian et al., 2009). Participants then underwent a cognitive battery consisting of several standardized examinations. The Trail-Making-Test (TMT) Part A measured visual scanning skills, attention, and processing speed, while TMT Part B further measured task-switching abilities and language-symbol manipulation (Gaudino et al., 1994). The Weschler Adult Intelligence Scale (WAIS) Digit Span test measured attention and short-term memory capacity when administered as a forward span, and executive function and working memory when administered as a backward span (Fink et al., 2014). Lastly, the WAIS Similarities test measured abstract thinking, concept formation, and verbal reasoning skills, while the WAIS Visual Puzzles test measured visual processing, attention, and fluid reasoning skills (Washington Center for Cognitive Therapy, 2015). Table 1 displays basic information about the participant groups, divided into the YA and OA cohorts.

Table 1. Participant Analysis | N = 38

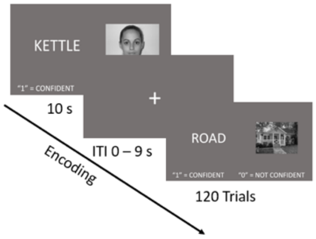
Sex (% Female) = 48%

Young Age (SD)		**Old Age (SD)**	
N	19	N	19
Sex (% F)	57%	Sex (% F)	35%
Age	25.8 (3.8)	Age	64.9 (4.0)
Age Range	21 - 35	Age Range	60 - 75
YoE	17.1 (2.2)	YoE	15.4 (2.1)

3.3 Encoding and retrieval

During the encoding phase, participants learned 120 word-picture pairs, where words were concrete nouns and pictures were either faces or houses. This process is visualized in Figure 1.

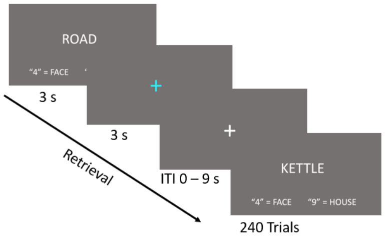
Figure 1. Word-picture pair encoding method



To increase task engagement and control for inconsistent memory strategies, participants were instructed to employ the 'interactive imagery' encoding strategy, where they describe a meaningful interaction between the word and the picture, and then rate their confidence that they will remember the pair in the future (Sahadevan et al., 2021). These encoding confidence (EC) ratings served as a measure of pre-test metamemory, also referred to as a prospective feeling-of-knowing (Chua et al., 2009). After 30-60 minutes, the participants entered retrieval, where they underwent a recall task while in a mock fMRI scanner. A mock scanner was employed in preparation for the next phase of this research, which will look to replicate this study using an fMRI scanner. The use of a mock scanner allows for the use of this paper's data as a pilot for the following experiment, because participants' memory performance is less likely to differ significantly when in a real fMRI scanner.

In the mock scanner, participants viewed previously encoded words and indicate with a button press whether the word had been paired with a face or a house, evaluating associative memory. This process is viewed in Figure 2.

Figure 2. Associative memory retrieval task



More specifically, their retrieval accuracy (whether their association is correct) and response time will be recorded. The order in which words were presented was randomized across all encoding and retrieval runs. Seven of the 38 participants did not complete all 240 retrieval trials, either because of technical malfunctions that began during the retrieval session or because of a delayed start time that led to conflicts with following mock scanner reservations.

3.4 Model construction

The first step was to build a hierarchical linear model (HLM) for the composite cohort, including all YA and OA participants. All HLMs were built in R using the 'lme4' package and tested throughout construction using the 'lmerTest' package. The data was first organized in Excel and then imported to R, where it was formatted appropriately. The model was built by first defining the simplest model – the intercept-only model – then introducing level-1 explanatory variables, and finally level-2 explanatory variables. The individual effect of each explanatory variable was evaluated in a stepwise manner using several tests to determine whether the variable was significantly useful in explaining variance or if it should be removed from the model. Statistical tests included the Likelihood Ratio Chi-Square Test (referred to as the Likelihood Ratio Test, or Deviance Test), Type 3 Analysis of Variance (ANOVA), and single-term deletions of random effects to generate ANOVA-like tables (Christensen, 2020).

The Likelihood Ratio Test yields a chi-square coefficient, which is the amount of deviance that is explained from one model to the next, and a p-value, which indicates whether the amount of explained deviance is significant given the additional degrees of freedom that were introduced. This test also yields an Akaike information criterion (AIC) and Bayesian information criterion (BIC). Both these values evaluate the quality of the model's fit; the BIC compares candidate models and assumes that the number of candidate models is fixed, whereas the AIC evaluates the predictive value of the current model to future unknown datasets (Vrieze, 2012). Lower AIC and BIC values are attributed to the model that is more likely to best fit the data.
The Type 3 ANOVA yields a table of the model's fixed effects as well as the interactions between fixed effects, as specified in a given model. Among the several calculated values are the F-value and p-value. When associated with a p-value that indicates significance, high F-values reflect a variable's significant contribution to the model. The single-term deletions of random effects method applies a Likelihood Ratio Test to yield a table with p-values, chi-square coefficients, and AIC terms. P-values indicate significance when they are less than the α-value ($\alpha = 0.05$).

Note that the terms yielded by these statistical tests were considered collectively, rather than making decisions based on the outcome of a single parameter. Further, if tests indicate no significance for a variable but the inclusion of that variable is deemed necessary based on literature-backed presumptions about its interaction with another significant variable, then the non-significant variable should be included (Snijders & Bosker, 2004).

All continuous variables except for response time are centered, which facilitates interpretation by equating the variable's mean value across all observations to zero. Level-1 variables are centered within-cluster (CWC), whereas level-2 variables are grand mean centered (GMC). For variable X with X_{avg} as its mean, X_{GMC} and X_{CWC} would be calculated as follows:

$$X_{GMC} = X - X_{avg\ across\ individuals}$$
$$X_{CWC} = X - X_{avg\ within\ given\ individual}$$

The effects of each variable are ultimately interpreted as a regression with other variables in the model. Centering allows the intercept of the variable's regression slope to be interpreted as the variable's average value across individuals (Snijders & Bosker, 2004). Response time is not centered because it is the outcome variable, and thus is not interpreted the way explanatory variables are.

Missing values were sporadically present in the data set for a variety of reasons, ranging from technical malfunctions to participant errors. Missing values were imputed using a multivariate imputation by chained equations method via the 'mice' R package. This allowed for the estimation of missing values based on observed values – rather than listwise deletion or available case analysis – to preserve the power of hypothesis tests (Snijders & Bosker, 2004).

The intercept-only model (Model 0) is a linear equation defining the i^{th} instance of response time for the j^{th} individual (RT_{ij}) as the mean RT across all individuals (the grand mean, γ_{00}) plus or minus the j^{th} individual's variability in RT (IIV-RT, u_{0j}) and the natural variability in RT between individuals (the residual effects, e_{ij}).

$$RT_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

This can be represented hierarchically, where terms pertaining to a single individual (i.e., age group, test scores) are clustered within the individual at level-2, while terms that generally describe instances at the trial level (i.e., RT, trial number) are at level-1. To model this, the coefficient β_{0j} represents level-2 for the j^{th} individual and is defined as the individual's variance (u_{0j}) from the grand mean RT (γ_{00}). RT_{ij} is then expressed as the intercept β_{0j} plus the residual variance at level-1, e_{ij}.

Model 0.

$$RT_{ij} = \beta_{0j} + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$

'Fixed' and 'random' effects can now be defined in the context of Model 0. An explanatory variable is said to have fixed effects if the intercept of its regression varies across individuals – indicating notable differences in the variable's average value (when the variable is centered). Fixed effects variables will thus have their own intercept term (γ_{xx}). An explanatory variable is said to have random effects if the slope of its regression varies across individuals – indicating notable differences in the variable's distribution. Random effects variables will thus have their variance term (u_{xx}). Variables can have both fixed and random effects, or just one of the two. If a variable does not have significant fixed or random effects, it is not a 'useful' parameter for the model and should be removed.

Trial-level level-1 variables were added to Model 0 in a stepwise manner, using the described statistical tests to determine whether the introduced variable should be modeled as a fixed effect, a random effect, both, or neither. As will be detailed later, it was determined that accuracy at retrieval (whether the subject correctly recalled the word-picture association) had fixed and random effects (Model 1):

Model 1.

$$RT_{ij} = \beta_{0j} + \beta_{1j}Accuracy_{ij} + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$

Note that the coefficient that denotes the effects of accuracy, β_{1j}, is defined by both an intercept and a variance term.

Retrieval trial number (out of a total of 240 trials) only had fixed effects (Model 2):

Model 2.

$$RT_{ij} = \beta_{0j} + \beta_{1j}Accuracy_{ij} + \beta_{2j}Trial_{ij} + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20}$$

Note that the coefficient that denotes the effects of trial number, β_{2j}, is defined by an intercept term but not a variance term.

At this point, level-2 variables were added to Model 2 in a stepwise manner, maintain the use of statistical tests to guide the decision-making process. Level-2 variables only have one value per individual, and thus have no within-individual variance. Accordingly, they can have fixed effects but will never have random effects. Furthermore, level-2 variables do not directly influence response time on a trial-by-trial basis. Rather, they 'moderate' the effects that trial-level (level-1) variables have on response time.

Determination of certain variables as irrelevant due to restricted range (MMSE, Boston Naming Test) and low moderating effects (YoE, TMT-A) resulted in the exclusion of certain potential level-2 variables.

Age Group (whether the participant is in the YA or OA cohort) had fixed effects and moderated the effects of trial number.

Model 2.1

$$RT_{ij} = \beta_{0j} + \beta_{1j}Accuracy_{ij} + \beta_{2j}Trial_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Age\ Group_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}Age\ Group_j$$

Digit Span test results had fixed effects with no moderator interactions:

Model 2.2

$$RT_{ij} = \beta_{0j} + \beta_{1j}Accuracy_{ij} + \beta_{2j}Trial_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Age\ Group_j + \gamma_{02}Digit\ Span_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}Age\ Group_j$$

TMT-B results had fixed effects and moderated the effects of trial number:

Model 2.3

$$RT_{ij} = \beta_{0j} + \beta_{1j}Accuracy_{ij} + \beta_{2j}Trial_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Age\ Group_j + \gamma_{02}Digit\ Span_j + \gamma_{03}TMT\ B_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}Age\ Group_j + \gamma_{23}TMT\ B$$

Finally, WAIS Similarities test results had fixed effects and also moderated the effects of trial number:

Model 2.4

$$RT_{ij} = \beta_{0j} + \beta_{1j}Accuracy_{ij} + \beta_{2j}Trial_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Age\ Group_j + \gamma_{02}Digit\ Span_j + \gamma_{03}TMT\ B_j + \gamma_{04}WAIS\ Sim_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}Age\ Group_j + \gamma_{23}TMT\ B_j + \gamma_{24}WAIS\ Sim_j$$

The finalized composite model (Model 2.4) was used to produce graphs and analyze the relationships between involved variables.

3.5 Age group analysis

To investigate differences between the YA and OA cohorts, I split the composite cohort into separate YA and OA data sets and applied each to Model 2.4 separately. The IIV-RT of each cohort was taken to be the level-2 variance term, given that individuals are represented in the model at level-2.

3.6 Split-cohort analyses

To further investigate this data set, I performed two sets of split-cohort analyses: one to investigate the effects of performance (in terms of retrieval accuracy) and one to investigate the effects of early versus late trials.

For the performance-wise split, I first calculated each participants' average accuracy (which I will call their 'performance'). Accuracy was encoded as 0 for a wrong association and 1 for a correct association. Correspondingly, performance was rated as a decimal between 0 and 1, where higher performers (those who were accurate more often) rated closer to one. I then found the average performance for each age cohort and then used this value to separate them into four sub-cohorts: a high performing YA cohort, a low performing YA cohort, a high performing OA cohort, and a low performing OA cohort. Each of these cohorts were then modeled using the same model that was constructed for the composite (all OA and all YA) data set (Model 2.4).

For the trial-wise split, I considered the first 120 trials to be 'early' and the final 120 trials to be 'late'. The result was another four sub-cohorts: an early trial YA cohort, a late trial YA cohort, an early trial OA cohort, and a late trial OA cohort. Again, each of these cohorts were modeled using Model 2.4. Not that in the trial-wise split participants are not being segregated within the age cohorts. Rather, each participant's data is being halved and analyzed separately. This is contrast to the performance-wise split, where participants were segregated within the age cohorts such that each sub-cohort consists of a distinct set of participants.

For all models, IIV-RT was taken to be represented by the level-2 variance term, because the individual is represented at level-2.

4. Results

4.1 Insights during composite model construction

The Likelihood Ratio Test was the primary test employed when including or excluding explanatory variables, given that it provides a variety of parameters that are relevant to the decision-making process. Table 2 provides the key parameters calculated through the Likelihood Ratio Test for all steps of model construction.

Table 2. Results of Likelihood Ratio Test Across Models

	Deviance	p-value	AIC	BIC
Model 0	21100	N/A	21106	21127
Model 1	20643	2.20E-16	20655	20698
Model 2	20470	2.20E-16	20484	20533
Model 2.1	20445	2.88E-06	20463	20526
Model 2.2	20440	2.41E-02	20460	20530
Model 2.3	20414	2.60E-06	20438	20522
Model 2.4	20402	2.71E-03	20430	20528

Overall, the parameters included in Model 2.4 decreased the unexplained deviance from Model 0 by 698 and decreased AIC and BIC scores by 676 and 599, respectively. The decrease in these scores indicates the development of a model that more accurately fits the given data set and accounts for previously unexplained variability in response time. Importantly, these values are paired with p-values consistently lower than the alpha value of 0.05, indicating that each model was significantly more 'fit' than its predecessor. Models that were excluded based on results from the Likelihood Ratio Test are not displayed, but the above parameters contributed to decisions to exclude models where encoding confidence, MMSE, WAIS Visual Puzzles, and other variables were included.

Table 2 also denotes variables that were particularly useful. The inclusion of accuracy as a fixed and random effect (Model 1), for example, decreased deviance by 457, AIC by 451, BIC by 429, and had a p-value of 2.20E-16 – a very significant test result. The addition of trial number as a fixed effect (Model 2), age group as a fixed effect and moderator of the trial number variable (Model 2.1), and TMT-B as a fixed effect and moderator of the trial number variable (Model 2.3) also had particularly significant p-values.

4.2 Insights from composite model

Table 3 displays the results of a Type 3 ANOVA test on the final model (Model 2.4). Note that the p-values are necessary as context for what size F-value is considered significant.

Table 3. ANOVA for Model 2.4

	F-value	p-value	<α
Accuracy	9.8752	3.50E-03	Yes
Trial Number	174.60	2.20E-16	Yes
Age Group	0.0003	9.87E-01	No
Digit Span	7.4635	9.96E-03	Yes
TMT-B	3.7514	6.27E-02	No
WAIS Sim.	2.2143	1.47E-01	No
Trial Num : AG	52.030	5.94E-13	Yes
Trial Num : TMT-B	13.130	2.92E-04	Yes
Trial Num : WAIS Sim.	9.2651	2.34E-03	Yes

This statistical test is useful because it details the significance and explanatory power of all the included variables in the context of each other, rather than in the sequential order that they were individually added (as shown in Table 2). The results shown in Table 3 also highlight the important point that not all added variables were significant for their direct explanatory power, but for their significant moderation of other variables (moderation is denoted in tables as *moderated variable*:*moderator variable*). For example, age group membership, TMT-B scores, and WAIS Similarities scores all have p-values that denote insignificant fixed effects (0.987, 0.063, and 0.147, respectively). However, all three variables significantly moderate the fixed effects of trial number (5.94E-13, 2.92E-04, and 2.34E-03, respectively).

Table 3 also demonstrates the comparatively significant contribution to the model of trial number. This variable's F-value (174.60) is more than triple the next highest (52.030) and is correspondingly accompanied by a highly significant p-value (2.20E-16).

Table 4. Single-Term Deletion for Model 2.4

	p-value	**Chi-Sq. Coef.**
Accuracy	2.20E-16	420.09

While the Type 3 ANOVA confirmed the significance of accuracy as a fixed effect (F-value = 9.8752; p-value = 3.50E-03), the single-term deletion of random effects method confirms its additional significance as a random effect. Chi-square coefficients are produced by the chi-square test, which is a part of the Likelihood Ratio Test. These coefficients represent the deviance that is explained by the inclusion of each random effect. The inclusion of accuracy's random effects reduced deviance by 420.09, a substantial portion of the total deviance explained (698). The produced p-value (2.20E-16) serves to emphasize the significance of accuracy's random effects. Together, the data presented in Tables 3 and 4 reveal that, though accuracy is significant in both its fixed and random effects, the latter are more critical to the model.

4.3 Age cohort analysis

The composite model was segregated into OA and YA models in order investigate in greater detail how age group affects IIV-RT and its implicated variables. Table 5 displays the results of a Type 3 ANOVA on Model 2.4 for the YA cohort. Note that, because all participants in this analysis belong to the same age group, the age group variable is redundant and thus not included.

Table 5. ANOVA for YA Model

	F-value	**p-value**	**<α**
Accuracy	7.0585	1.78E-02	Yes
Trial Number	118.14	2.20E-16	Yes
Digit Span	3.8966	0.06668	No
TMT-B	2.8490	0.11363	No
WAIS Sim.	0.0002	0.98892	No
Trial Num : TMT-B	3.4070	0.06499	No
Trial Num : WAIS Sim.	1.7713	0.18329	No

There are notable differences between Tables 3 and 5. Namely, only trial-level variables were significant in determining response time (Accuracy F-value = 7.0585, p-value = 0.0178; Trial Num F-value = 118.14, p-value = 2.20E-16). None of the level-2 cognitive assessment scores significantly contributed to the model. None of these level-2 variables significantly moderated the effects at level-1, either.

As shown in Table 6, a single-term deletion of random effects method reveals that, as was the case with the composite model, the random effects of accuracy (p-value = 2.20E-16) are more significant than its fixed effects (p-value = 1.78E-02), although both are significant.

Table 6. Single-Term Deletion for YA Model

	p-value	**Chi-Sq. Coef.**
Accuracy	2.20E-16	143.75

Table 7 displays the results of a Type 3 ANOVA on Model 2.4 for the OA cohort.

Table 7. ANOVA for OA Model

	F-value	**p-value**	**<α**
Accuracy	2.8617	0.10774	No
Trial Number	2.2653	0.13238	No
Digit Span	2.1721	0.16149	No
TMT-B	2.5323	0.13343	No
WAIS Sim.	7.3785	0.01603	Yes
Trial Num : TMT-B	9.7660	0.00179	Yes
Trial Num : WAIS Sim.	10.087	0.00150	Yes

The results of this table are almost complementary to those of Table 5. Here we see that no trial-level variables were significant in determining response time. However, WAIS Sim. scores significantly contributed to the model (F-value = 7.3785; p-value = 0.01603) and moderated the effects of trial number (F-value = 10.087; p-value = 0.00150). TMT-B scores also significantly moderated the effects of trial number (F-value =9.7660; p-value = 0.00179).

4.4 Performance-wise split-cohort analysis

The average performance (ranging from 0 for 100% inaccurate to 1 for 100% accurate) was 0.773 for YA participants and 0.670 OA participants. The ratio of below-average to above-average performers was 7-12 for the YA cohort and 10-9 for the OA cohort. For both OA and YA participants, low performers (LP) had greater IIV-RT (LP OA = 0.461 sec.; LP YA = 0.602 sec.) than high performers (HP; HP OA = 0.343 sec.; HP YA = 0.311 sec.). For reference, the average response times were equal for both cohorts (2.33 sec.). See appendix for scatter plots of response time by trial number for each individual in the LP OA (Supp. Fig. 1), HP OA (Supp. Fig. 2), LP YA (Supp. Fig. 3), and HP YA (Supp. Fig. 4) sub-cohorts.

Table 8 displays the results of a Type 3 ANOVA on the LP YA and HP YA models. The significance of trial number in determining response time seen in the general YA cohort model (Table 5) persists across both LP (F-value = 67.36; p-value = 5.26E-16) and HP (F-value = 40.11; p-value = 2.82E-10) YA participants. Accuracy was a significant determinant of response time for HP YA participants (F-value = 13.81; p-value = 0.007), as were Digit Span scores (F-value = 8.414; p-value = 0.019). Neither of these variables were significant for LP YA participants. Notably, the lack of between-level interactions observed in the general YA cohort model persists across both LP and HP YA participants.

Table 8. ANOVA for LP and HP YA Models

	F-value	**p-value**	**<α**
----------------*Low Performers (LP)* ----------------			
Accuracy	0.015	0.906	No
Trial Number	67.36	5.26E-16	Yes
Digit Span	0.115	0.757	No
TMT-B	1.144	0.363	No
WAIS Sim.	0.195	0.689	No
Trial Num : TMT-B	0.402	0.526	No
Trial Num : WAIS Sim.	3.224	0.073	No
----------------*High Performers (HP)*----------------			
Accuracy	13.81	0.007	Yes
Trial Number	40.11	2.82E-10	Yes
Digit Span	8.414	0.019	Yes
TMT-B	5.204	0.051	No
WAIS Sim.	0.232	0.643	No
Trial Num : TMT-B	0.469	0.493	No
Trial Num : WAIS Sim.	0.315	0.575	No

Table 9 displays the results of a Type 3 ANOVA on the LP OA and HP OA models.

Table 9. ANOVA for LP and HP OA Models

	F-value	**p-value**	**<α**
----------------*Low Performers (LP)* ----------------			
Accuracy	0.047	0.833	No
Trial Number	0.184	0.668	No
Digit Span	0.282	0.614	No
TMT-B	0.219	0.656	No
WAIS Sim.	0.363	0.568	No
Trial Num : TMT-B	8.354	0.004	Yes
Trial Num : WAIS Sim.	0.755	0.385	No
----------------*High Performers (HP)*----------------			
Accuracy	62.21	6.75E-05	Yes
Trial Number	4.037	0.045	Yes
Digit Span	0.093	0.772	No
TMT-B	3.365	0.124	No
WAIS Sim.	0.251	0.639	No
Trial Num : TMT-B	4.554	0.033	Yes
Trial Num : WAIS Sim.	8.895	0.003	Yes

As was the case for the general OA cohort, trial number is not a significant determinant of response time for LP OA participants. However, it is significant for HP OA participants (F-value = 4.037; p-value = 0.045). Similarly, accuracy is not a significant determinant of response time for LP OA participants, but it is very significant for

HP OA participants (F-value = 62.21; p-value = 6.75E-05).

There is also a continuation of the between-level interactions observed in the general OA cohort (Table 7). TMT-B scores significantly moderated the effects of trial number in both LP (F-value = 8.354; p-value = 0.004) and HP participants (F-value = 4.554; p-value = 0.033). WAIS Sim. scores significantly moderated the effects of HP OA participants only (F-value = 8.895; p-value = 0.003).

4.5 Trial-wise split-cohort analysis

Within the YA cohort, early trials were associated with greater IIV-RT (0.375 sec.) than late trials (0.218 sec.). With the OA cohort, early trials were associated with lesser IIV-RT (0.289 sec.) than late trials (0.898 sec.). See appendix for scatter plots of response time by trial number for each individual in the early trial OA (Supp. Fig. 5), late trial OA (Supp. Fig. 6), early trial YA (Supp. Fig. 7), and late trial YA (Supp. Fig. 8) sub-cohorts.

Table 10 displays the results of a Type 3 ANOVA on the early trial (ET) YA and late trial (LT) YA models.

Table 10. ANOVA for ET and LT YA Models

	F-value	p-value	<α
---------------------*Early Trial (ET)*--------------------			
Accuracy	7.002	0.018	Yes
Trial Number	7.331	0.007	Yes
Digit Span	2.417	0.14	No
TMT-B	1.501	0.238	No
WAIS Sim.	0.596	0.451	No
Trial Num : TMT-B	0.172	0.678	No
Trial Num : WAIS Sim.	8.477	0.004	Yes
---------------------*Late Trial (LT)*--------------------			
Accuracy	8.177	0.014	Yes
Trial Number	3.731	0.054	No
Digit Span	5.306	0.043	Yes
TMT-B	3.859	0.074	No
WAIS Sim.	0.164	0.693	No
Trial Num : TMT-B	1.592	0.207	No
Trial Num : WAIS Sim.	1.837	0.175	No

The significance of trial number in determining response time for YA participants reported in Tables 5 and 8 is again seen in early trials (F-value = 7.331; p-value = 0.007), but not late trials. Similarly, WAIS Sim. scores significantly moderated the effects of trial number in early (F-value = 8.477; p-value = 0.004) but not late trials. Accuracy, however, significantly determined response times across both early (F-value = 7.002;, p-value = 0.018) and late (F-value = 8.177; p-value = 0.014) trials.

Table 11 displays the results of a Type 3 ANOVA on the ET OA and LT OA models.

Table 11. ANOVA for ET and LT OA Models

	F-value	p-value	<α
--------------------*Early Trial (ET)*--------------------			
Accuracy	0.264	0.123	No
Trial Number	0.025	0.874	No
Digit Span	2.449	0.139	No
TMT-B	2.435	0.139	No
WAIS Sim.	11.77	0.003	Yes
Trial Num : TMT-B	2.382	0.123	No
Trial Num : WAIS Sim.	13.59	2.33E-04	Yes
--------------------*Late Trial (LT)*--------------------			
Accuracy	24.42	1.98E-04	Yes
Trial Number	4.645	0.031	Yes
Digit Span	3.631	0.076	No
TMT-B	2.456	0.138	No
WAIS Sim.	2.605	0.126	No
Trial Num : TMT-B	0.192	0.661	No
Trial Num : WAIS Sim.	0.08	0.777	No

For OA participants, trial number significantly determined response time in late trials (F-value = 4.645; p-value = 0.031) but not in early trials. Similarly, accuracy very significantly determined response time in late trials (F-value = 22.42; p-value = 1.98E-04), but not in early trials. Conversely, WAIS Sim. scores very significantly moderated the effects of trial number in early trials (F-value = 13.59; p-value = 2.33E-04), but not late trials.

5. Discussion

The results reported in this paper provide further support for the relation between IIV-RT and neurocognitive health. More specifically, these results may be suggestive of a difference in cognitive strategy between age groups. While YA participants seem to approach the associative memory retrieval paradigm in a highly task-dependent manner, OA participants seem to rely more on their cognitive reserve. This is in line with existing evidence that OA individuals have a greater CR than YA individuals (Zihl et al., 2014). These differences are evident in several facets of the presented analyses, which I will now discuss.

5.1 Age group

The evidence that contributes to this hypothesis in the most straightforward manner is that presented in Tables 5-7. Table 5 demonstrates that, when the YA cohort was modeled independently of the OA cohort, response time was significantly determined by accuracy and trial number and nothing else. Accuracy and trial number are the only two variables in the presented models that correspond to the task that, at any given trial, a participant is engaged with. That is, while level-2 variables that correspond with a participant's age or neurocognitive health are static across all 240 trials and thus not task-dependent, accuracy and trial number change as the 240 trials progress.

In contrast, when the OA cohort was modeled independently of the YA cohort, neither of the task-dependent variables significantly determined response time. Instead, WAIS Sim. and TMT-B scores' interactions with trial number were significant. Remember that HLMs are structured such that level-2 variables cannot directly determine the outcome variable (response time). Instead, they exert their influence on the outcome variable through the moderation of level-1 variables, which is possible even when such variables are not themselves significant.

These results directly suggest that during the recall process, after accounting for variables that are task-dependent and those that pertain to the cognitive reserve, YA participants use strategies that implicate task-dependent variables whereas OA participants use strategies that utilize their cognitive reserve.

5.2 Trial number

Trial number, a variable that essentially represents the 'location' of a participant across the 240 retrieval trials, proved to be a highly significant variable (Table 3). Although this can be interpreted as a mere demonstration that response times are not static across trials, it also indicates that any given response time is significantly determined by its context. In other words, in attempting to predict what a response time will be, it is useful to consider whether the response time is early or late in the 240 trials. This, in and of itself, is a critical takeaway because it demonstrates that there is value to the HLM approach. Without this approach, a response time's location is lost in the process of aggregation, where all response times are collapsed into a single data point.

It is also noteworthy that all significant moderating interactions involved trial number. No level-2 variables significantly moderated the effects of retrieval accuracy. However, all except for digit span significantly moderated trial number. The fixed effects correlation matrix from the composite model (see Supplemental Table 1) shows that people with lower TMT-B scores (i.e., people that completed the task faster and thus performed better) have response times that are influenced by response time to a lesser extent. Similarly, people with higher (i.e., better) WAIS Sim. scores have response times less susceptible to trial number. Together, these two interactions provide evidence that individuals that performed better on neurocognitive assessments (and are assumed to be more neurocognitively healthy) have more stable response time patterns.

5.3 Early-to-late trial adaptations

The role of trial number was investigated further by performing a trial-wise split-cohort analysis. This analysis revealed a notable number of differences in relevant variables between early and late trials. It also yielded evidence for the previously mentioned concept of task-dependence versus cognitive reserve.

Table 10 shows that the response times of YA participants were significantly determined by accuracy and trial number in early trials, reflecting a high degree of task-dependence in the production of response times. No level-2 variables were themselves significant for the early trial YA model, although higher WAIS Sim. scores did moderate trial number such that it had a lesser effect on response time (see Supplemental Table 2). In contrast, the response times of OA participants in early trials were not determined by task-dependent variables (Table 11). Instead, WAIS Sim. scores were the only significant variable. These results suggest that OA participants depend less on task-dependent factors as they produce response times, and more on static measures of their cognitive ability – contributing to the hypothesis that they resort to cognitive reserve more than YA participants.

As participants progressed through the late trials, however, the significance of variables changed. Task-dependent variables became more prominent in the OA cohort, with accuracy and trial number both becoming more significance, whilst WAIS Sim. scores lost their significance. Although trial number lost its significance in the YA cohort's late trials, accuracy remained an important factor. These changes suggest that factors associated with later trials, such as fatigue, cause participants to adapt towards more task-dependent strategies such that a trial's context and accuracy affect response time to a greater extent.

An individual's ability to rely on their cognitive reserve for a task is as dependent on the cognitive load of the task as it is on the individual's cognitive health (Montemurro et al., 2019). If we assume that later trials are associated with a greater cognitive load – due to increased recall difficulty, distractions, and fatigue – we might expect that OA participants' dependence on their cognitive reserve would leave them vulnerable to late-trial performance that resemble greater cognitive deficits. Indeed, the OA cohort's IIV-RT increased from 0.289 seconds in early trials to 0.898 in late trials. Meanwhile, the YA cohort – presumably less dependent on cognitive reserve – decreased their IIV-RT from 0.375 seconds to 0.218 seconds between early and late trials.

This finding has several implications. First, we must consider that, because the cognitive load of a task and the cognitive health of an individual affect cognitive reserve in similar ways, studies must be weary that differences in cognitive load represent a potential confound when attempting to measure cognitive deficits. In this study, it is possible that the increase in cognitive load from early to late trials exacerbates the increase in OA IIV-RT that may otherwise have been fully attributed to OA participants' greater cognitive deficits. On the same note, however, if we accept that differences in cognitive load interact with cognitive reserve similarly to differences in cognitive deficits, then these results may be taken as preliminary evidence that IIV-RT can successfully differentiate degrees of cognitive load (and thus cognitive deficits) in cognitive reserve-dependent tasks. Future experiments may build on this by engaging the cognitive reserve while controlling for cognitive load and seeing if IIV-RT successfully differentiates degrees of cognitive deficits.

5.4 Differences between low and high performers

The other significant level-1 variable was retrieval accuracy. Although it is incorporated in a different manner than the level-2 cognitive test scores, retrieval accuracy is itself also a measure of neurocognitive health. Ultimately, one would expect individuals that remember the encoded word-picture pairs better to have healthier underlying neural substrates, and thus lower IIV-RT (Walhovd & Fjell, 2007).

The performance-wise split-cohort analysis found just this (Tables 8 and 9). Across both YA and OA cohorts, those who recalled the word-picture pairs with above-average accuracy had lower IIV-RT than those who had below-average accuracy. This serves as another demonstration that IIV-RT is a useful measure of cognitive ability and can successfully differentiate between high and low performers of a cognitive task.

The performance-wise split-cohort analysis also demonstrated that, across YA and OA cohorts, accuracy only significantly determined response time for high performers. That is, whether a given recall was correct or incorrect was not a relevant factor in determining the response time of that recall for low-performing individuals, regardless of age. This suggests that high performers engage active retrieval strategies that implicate conscious, knowledge-based processes that are affected by success (culmination in a correct response). In contrast, low performers may engage in a retrieval strategy that is simpler and less associated with the access of encoded knowledge, where the ultimate success of retrieval is less relevant.

Yonelinas and Jacoby (1994) have proposed a model for distinguishing the neural processes of recall based on *recollection* versus *familiarity*, called the 'two-factor theory of recognition memory'. They argue that recollection is a conscious process that involves knowledge-based memory (i.e., descriptions), whereas familiarity is a non-conscious process involving simple memory (i.e., discriminations) without qualitatively descriptive information.

Importantly, recollection is associated with less mistakes at recall (Jacoby, 1991). It would thus be consistent that low performers engage in the more error-prone process of familiarity-based recall, whilst high performers engage in the more accurate recollection-based recall. Furthermore, whereas the only significant level-2 variable across the two low performer models was TMT-B scores' moderation of trial number in the OA cohort, there was demonstrated significance of Digit Span scores for YA high performers and significant moderations of trial number by TMT-B and WAIS Sim. scores for OA high performers. This increased significance of cognitive assessment scores is suggestive of greater involvement of cognitive processes in high performers, relative to low performers. This is consistent with Yonelinas & Jacoby's proposition that recollection is a more cognitively complex process than familiarity.

5.5 Conclusion

This paper has provided evidence that lower IIV-RT is significantly related to performance on cognitive assessments such as TMT-B and WAIS Similarities that reflects greater neurocognitive health. It has also demonstrated that individuals that recall previously encoded word-picture pairs with above-average accuracy have lower IIV-RT than below-average counterparts within their age group.

By segregating the initial HLM by age, the results of this study have suggested that, during an associative memory recall task, YA individuals' response times are determined more by task-dependent variables than variables pertaining to their cognitive reserve, whereas the opposite is true for OA individuals. Moreover, a further segregation of the age-specific HLMs into early versus late trials revealed that OA participants' suggested dependence on cognitive reserve leaves them vulnerable to the effects of an increased cognitive load, resulting in IIV-RT scores that resemble that of greater cognitive deficits.

Finally, a segregation of age specific HLMs into high versus low performers suggested that high performing individuals utilize a more cognitively demanding recall strategy called recollection, whereas low performing individuals utilize a less cognitively demanding recall strategy called familiarity.

5.6 Future Directions

There are elements of the reported data that remain to be explained, either through further theoretical consideration or experimental investigation. For example, the YA cohort surprisingly had higher IIV-RT in the early cohorts than in the late cohorts – contradicting the expected effects of fatigue. Also, to better understand differences in retrieval strategies between high and low performers, future instantiations of this recall paradigm should prompt participants to rate how confident they are in their response after each retrieval trial. These confidence ratings will provide insight into whether a given response was intentionally given based on conscious recollection or non-conscious familiarity.

The logical next step for this project is to move past the dependence on neurocognitive assessments, which are ultimately secondary measures of neural health, and collect direct measures of neural health via functional magnetic resonance imaging (fMRI). Neurocognitive assessment scores should not be abandoned but combined with fMRI data in a HLM of similar structure for a more comprehensive investigation of the relation between IIV-RT and neural health.

Mueller et al. (2013) reported a correlation between brain folding pattern and functional variability – where brain folding pattern is defined by sulcal depth variability. They replicated existing findings that sulcal depth variability was highest in the lateral frontal and temporoparietal regions. The lateral frontal region contains the inferior frontal gyrus, which is involved in language processing. The left and right lateral temporoparietal regions are implicated in language comprehension and the ventral attention network, respectively. These are cognitive processes that are critical to many memory paradigms, including the one engaged in the present study. Thus, these ROIs should be specifically investigated in future projects, along with whole-brain analyses to identify any potential ROIs that are unique to this experimental paradigm.

White matter integrity is evidently implicated in the neural substrates of IIV-RT, and it should certainly be investigated in future studies using diffusion tensor imaging techniques. However, it is also important to investigate the relationship between IIV and task-related brain function. MacDonald et al. (2008) account for task-related brain function by first performing whole-brain fMRI analysis to identify regions of interest (ROIs) that are (statistically) significantly activated during individual trials, then investigating correlations between the degree of activation of these ROIs and degree of IIV-RT. They reported several ROIs, including the supramarginal gyrus – which again implicates folding patterns – and the hippocampal complex.

This study provides an ideal steppingstone towards more direct measurements of the neural substrates of IIV-RT, which represents a crucial step towards using IIV-RT as a diagnostic measure for more sensitive detection of early-stage cognitive impairment.

6. References

n.a. (2021). 2021 Alzheimer's disease facts and figures. *Alzheimer's & Dementia,* 17(3):327-406. DOI: 10.1002/alz.12328

Alves, L., Cardoso, S., Maroco, J., de Mendonca, A., Guerreiro, M., & Silva, D. (2018). Neuropsychological Predictors of Long-Term (10 Years) Mild Cognitive Impairment Stability. *Journal of Alzheimer's Disease,* 62(4):1703-11. DOI: 10.3233/JAD-171034

Aprahamian, I., Martinelli, J.E., Neri, A.L., & Yassuda, M.S. (2009). "The Clock Drawing Test: A review of its accuracy in screening for dementia." Dementia & Neuropsychology 3(2):74-81. DOI:10.1590/S1980-57642009DN30200002

Bauer, D.J., Gottfredson, N.C., Dean, D., & Zucker, R.A. (2013). Analyzing Repeated Measures Data on Individuals Nested within Groups: Accounting for Dynamic Group Effects. *Psychological Methods,* 18(1):1-14. doi: 10.1037/a0030639

Backman, L., Nyberg, L., Lindenberger, U., Li, S., & Farde, L. (2006). The correlative triad among aging, dopamine, and cognition: Current status and future prospects. *Neuroscience & Biobehavioral Reviews*, 30(6):791-807. https://doi.org/10.1016/j.neubiorev.2006.06.005

Christ, B.U., Combrinck, M.I., & Thomas, K.G.F. (2018). Both Reaction Time and Accuracy Measures of Intraindividual Variability Predict Cognitive Performance in Alzheimer's Disease. *Frontiers in Human Neuroscience,* 12. doi:10.3389/fnhum.2018.00124.

Chua, E.F., Schacter, D.L., & Sperling, R.A. (2009). “Neural correlates of memory: a comparison of feeling-of-knowing and retrospective confidence judgements.” Journal of Cognitive Neuroscience 21(9):1751-1765. DOI:10.1162/jocn.2009.21123

Fink, H.A., Hemmy, L.S., MacDonald, R., Carlyle M.H., Olson, C.M., Dysken, M.W., McCarten, J.R., Kane, R.L., Rutks, I.R., Ouellette, J., & Wilt, T.J. (2014). Cognitive Outcomes After Cardiovascular Procedures in Older Adults: A Systematic Review. Agency for Healthcare Research and Quality.

Gaudino, E.A., Geisler, M.W., & Squires, N.K. (1994). “Construct validity in the trail making test: What makes part B harder?” Journal of Clinical and Experimental Neuropsychology 17(4):529-535. DOI:10.1080/01688639508405143

Hebert, L.E., Weuve, J., Scherr, P.A., & Evans, D.A. (2013). Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. *Neurology,* 80(19):1778-83. DOI: 10.1212/WNL.0b013e31828726f5

Hox, J.J. (2010). Multilevel Analysis: Techniques and Applications (2nd ed.). Routledge.

Li, R., Yin, S., Zhu, X., Ren, W., Yu, J., Wang, P., Zeng, Z., Niu, Y.N., Huang, X., & Li, J. (2017). Linking Inter-Individual Variability in Functional Brain Connectivity to Cognitive Ability in Elderly Individuals. *Frontiers in Aging Neuroscience*, 9. doi: 10.3389/fnagi.2017.00385. https://doi.org/10.3389/fnagi.2017.00385

Li, S.C. & Lindenberger, U. (1999). Cross-level unification: A computational exploration of the link between deterioration of neurotransmitter systems and dedifferentiation of cognitive abilities in old age. *Cognitive Neuroscience of Memory,* 103-146.

Li, S.C., Lindenberger, U., & Sikstrom, S. (2001). Aging cognition: from neuromodulation to representation. *Trends in Cognitive Sciences*, 5(11):479-486. DOI:10.1016/s1364-6613(00)01769-1

Linn, R.T., Wolf, P.A., Bachman, D.L., Knoefel, J.E., Cobb, J.L., Belanger, A.J., Kaplan, E.F., & D’Agostino, R.B. (1995). The ‘preclinical phase’ of probable Alzheimer’s disease. A 13-year prospective study of the Framingham cohort. *Archives of Neurology*, 52(5):485-90. DOI: 10.1001/archneur.1995.00540290075020

Locascio, J.J., Growdon, J.H., & Corkin. S. (1995). Cognitive test performance in detecting, staging, and tracking Alzheimer’s disease. *Archives of Neruology*, 52(11):1087-99. DOI: 10.1001/archneur.1995.00540350081020

MacDonald, S.W.S, Nyberg, L., Sandblom, J., Fischer, H., & Backman, L. (2008). Increased Response-time Variability is Associated with Reduced Inferior Parietal Activation during Episodic Recognition in Aging. *Journal of Cognitive Neuroscience*, 20(5): 779-786. doi:10.1162/jocn.2008.20502

Mella, N., de Ribaupierre, S., Eagleson, R., & de Ribaupierre, A. (2013). Cognitive Intraindividual Variability and White Matter Integrity in Aging. *The Scientific World Journal*, doi:10.1155/2013/350623

Montemurro, S., Mondini, S., Crovace, C., & Jarema, G. (2019). Cognitive Reserve and Its Effect in Older Adults on Retrieval of Proper Names, Logo Names and Common Nouns. *Frontiers in Communication,* 4(14):1-12.

https://doi.org/10.3389/fcomm.2019.00014

Mueller, S., Wang, D., Fox, M. D., Yeo, B. T. T., Sepulcre, J., Sabuncu, M.R., Shafee, R., Lu, J., & Liu, H. (2013) Individual Variability in Functional Connectivity Architecture of the Human Brain. *Neuron,* 77(3):586-95. https://doi.org/10.1016/j.neuron.2012.12.028

Opar, A. (2010). Hope builds for earlier detection of Alzheimer's disease: new recommendations to change the diagnostic criteria for Alzheimer's disease and recent advances in biomarker development may allow earlier diagnosis of this disease, potentially providing key tools for drug development. *Nature Reviews Drug Discovery,* 9(8): 579-81. DOI: https://doi.org/10.1038/nrd3237

Pasquier, F. (1998). Early diagnosis of dementia: neuropsychology. *Journal of Neurology,* 246:6-15. DOI: https://doi.org/10.1007/s004150050299

Sahadevan, S.S., Chen, Y.Y., & Caplan, J.B. (2021). Imagery-based strategies for memory for associations. *Memory,* 29(10):1275-95. https://doi.org/10.1080/09658211.2021.1978095

Snijders, T.A.B. & Bosker, R.J. (2004). Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling (2nd ed.). Sage.

Snyder, P.J., Kahle-Wrobleski, K., Brannan, S., Miller, D.S., Schindler, R.J., Desanti, S., Ryan, M.J., Morrison, G., Grundman, M., Chandler, J., Caselli, R. J., Isaac, M., Bain, L., & Carrillo, M.C. (2014). Assessing cognition and function in Alzheimer's disease clinical trials: do we have the right tools? *Alzheimer's & Dementia*, 10(6):853–60. DOI:10.1016/j.jalz.2014.07.158

Stefanis, N.C., van Os, J., Avramopoulos, D., Smyrnis, N., Evdokimidis, I., Stefanis, C.N. (2005). Effect of COMT Val^{158}Met Polymorphism on the Continuous Performance Test, Identical Pairs Version: Tuning Rather Than Improving Performance. *The American Journal of Psychiatry*, 162(9):1752-4. https://doi.org/10.1176/appi.ajp.162.9.1752

Tombaugh, T.N. & McIntyre, N.J. (1992). "The mini-mental state examination: a comprehensive review." Journal of the American Geriatrics Society 40(9):922-935. DOI:10.1111/j.1532-5415.1992.tb01992.x

Walhovd, K.B. & Fjell, A.M. (2007). White matter volume predicts reaction time instability. *Neuropsychologia,* 45(10):2277-84. https://doi.org/10.1016/j.neuropsychologia.2007.02.022

Washington Center for Cognitive Therapy (2015). "Description of WAIS-IV Subtests." Weschler Adult Intelligence Scale-IV. Retrieved from: http://www.washingtoncenterforcognitivetherap y.com/wp-content/uploads/2015/01/greenwood_description -wais-1.pdf

Wheeler, M.E., Woo, S.G., Ansel, T., Tremel, J.J., Collier, A.L., Velanova, K., Ploran, E.J., & Yang, T. (2015). The Strength of Gradually Accruing Probabilistic Evidence Modulates Brain Activity During a Categorical Decision. *Journal of Cognitive Neuroscience,* 27(4): 705-19. doi:10.1162/jcon_a_00739

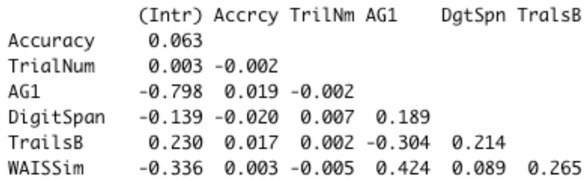
Woltman, H., Feldstain, A., MacKay, J.C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods in Psychology*, 8(1): 52- 69. doi:10.20982/tqmp.08.1.p052

Yiannopoulou, K.G. & Papageorgiou S.G. (2020). Current and Future Treatments in Alzheimer Disease: An Update. *Journal of Central Nervous System Disease*. DOI: 10.1177/1179573520907397

Zihl, J., Fink, T., Pargent, F., Ziegler, M., & Buhner, M. (2014). Cognitive Reserve in Young and Old Healthy Subjects: Differences and Similarities in a Testing-the-Limits Paradigm with DSST. *PLoS ONE,* 9(1):e84590. doi:10.1371/journal.pone.0084590

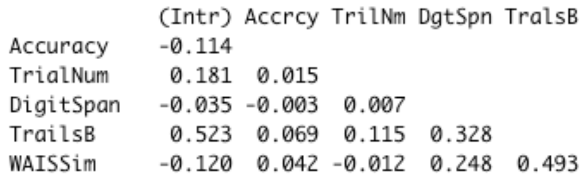
7. Appendix

Supplemental Table 1: Fixed Effects Correlation Matrix for Composite Model



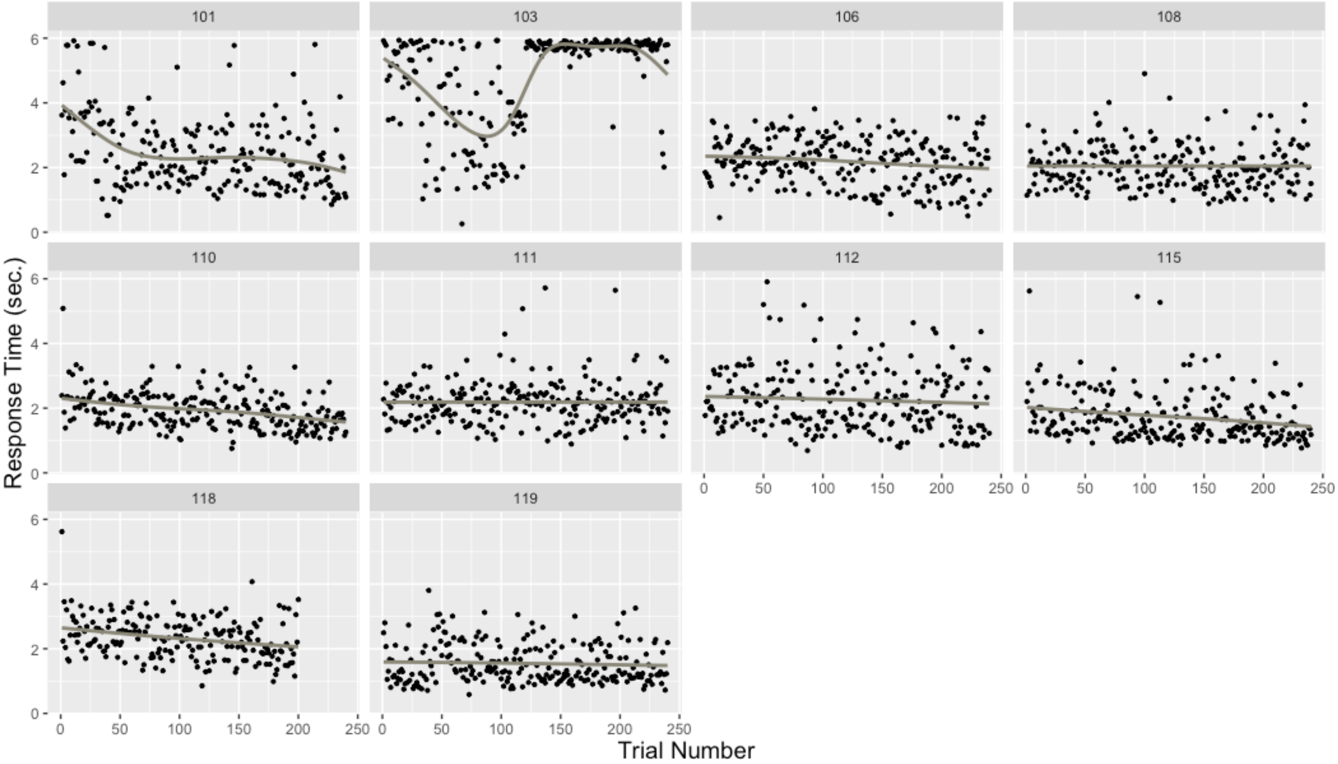
```
          (Intr) Accrcy TrilNm AG1    DgtSpn TralsB
Accuracy   0.063
TrialNum   0.003 -0.002
AG1       -0.798  0.019 -0.002
DigitSpan -0.139 -0.020  0.007  0.189
TrailsB    0.230  0.017  0.002 -0.304  0.214
WAISSim   -0.336  0.003 -0.005  0.424  0.089  0.265
```

Supplemental Table 2: Fixed Effects Correlation Matrix for ET YA Model

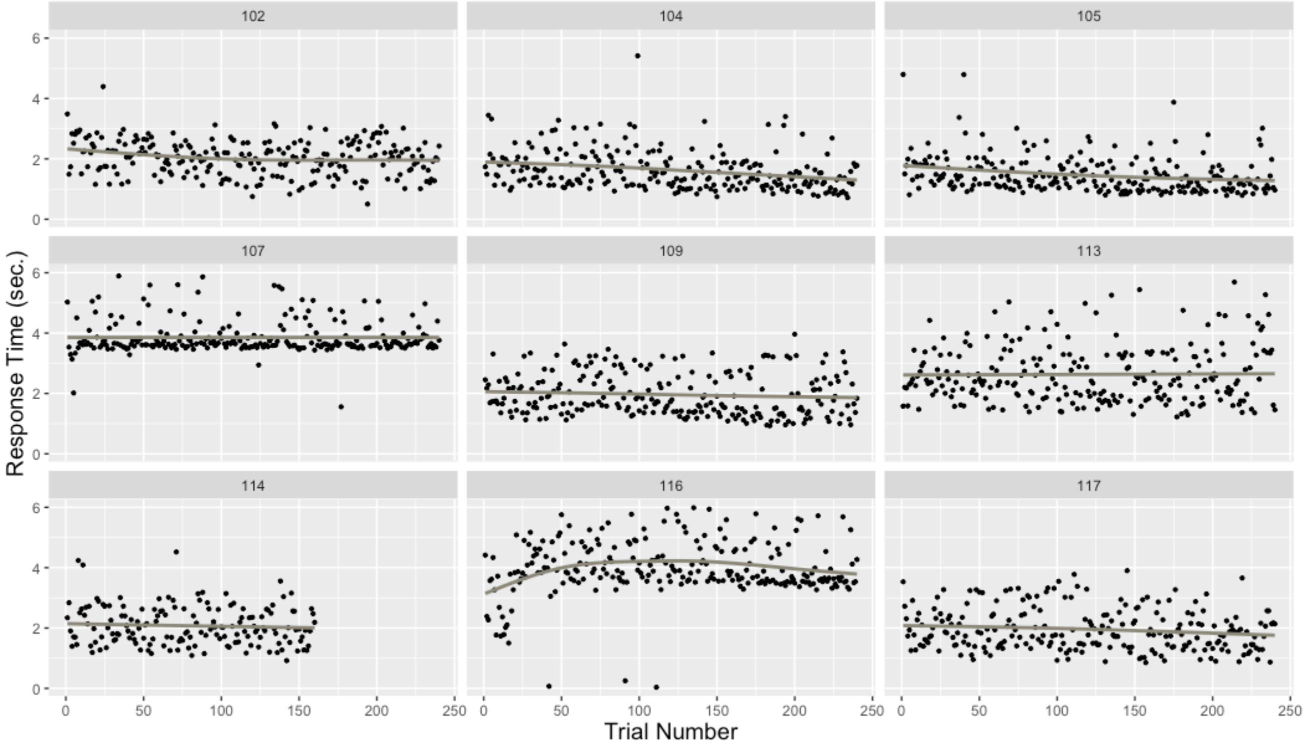


```
          (Intr) Accrcy TrilNm DgtSpn TralsB
Accuracy  -0.114
TrialNum   0.181  0.015
DigitSpan -0.035 -0.003  0.007
TrailsB    0.523  0.069  0.115  0.328
WAISSim   -0.120  0.042 -0.012  0.248  0.493
```

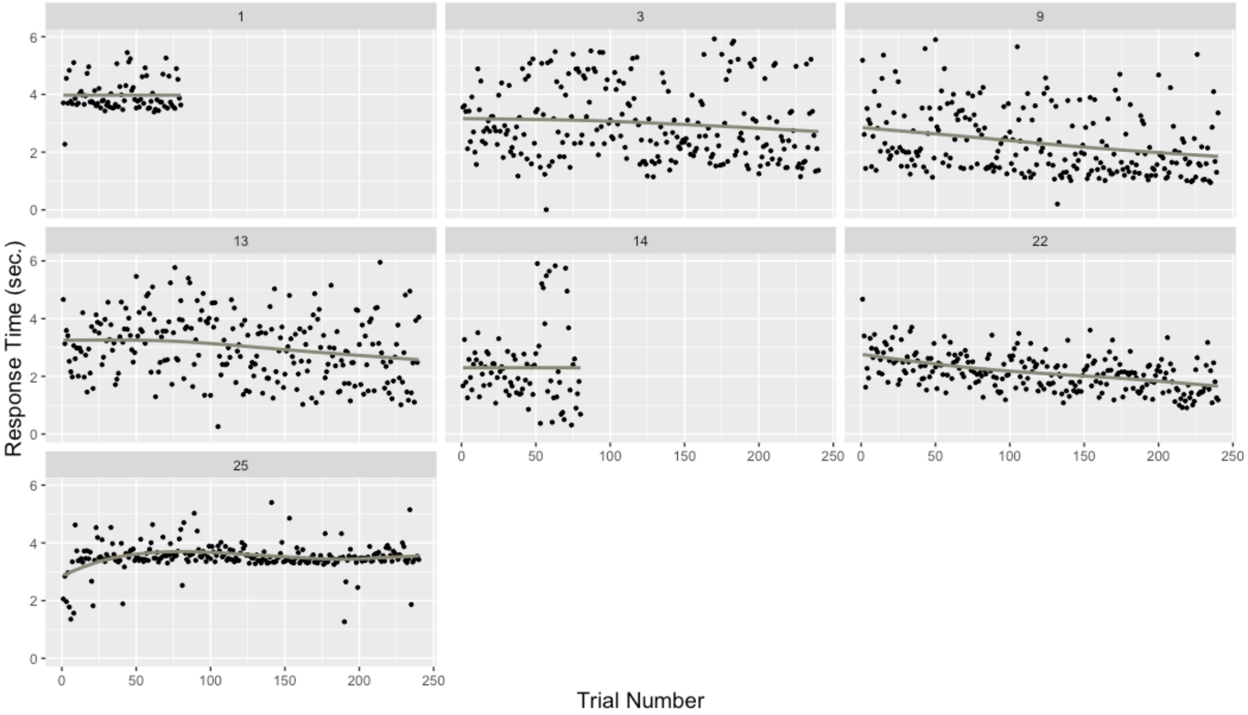
Supplemental Figure 1: Trial Number by Response Time for Low Performing OA Individuals



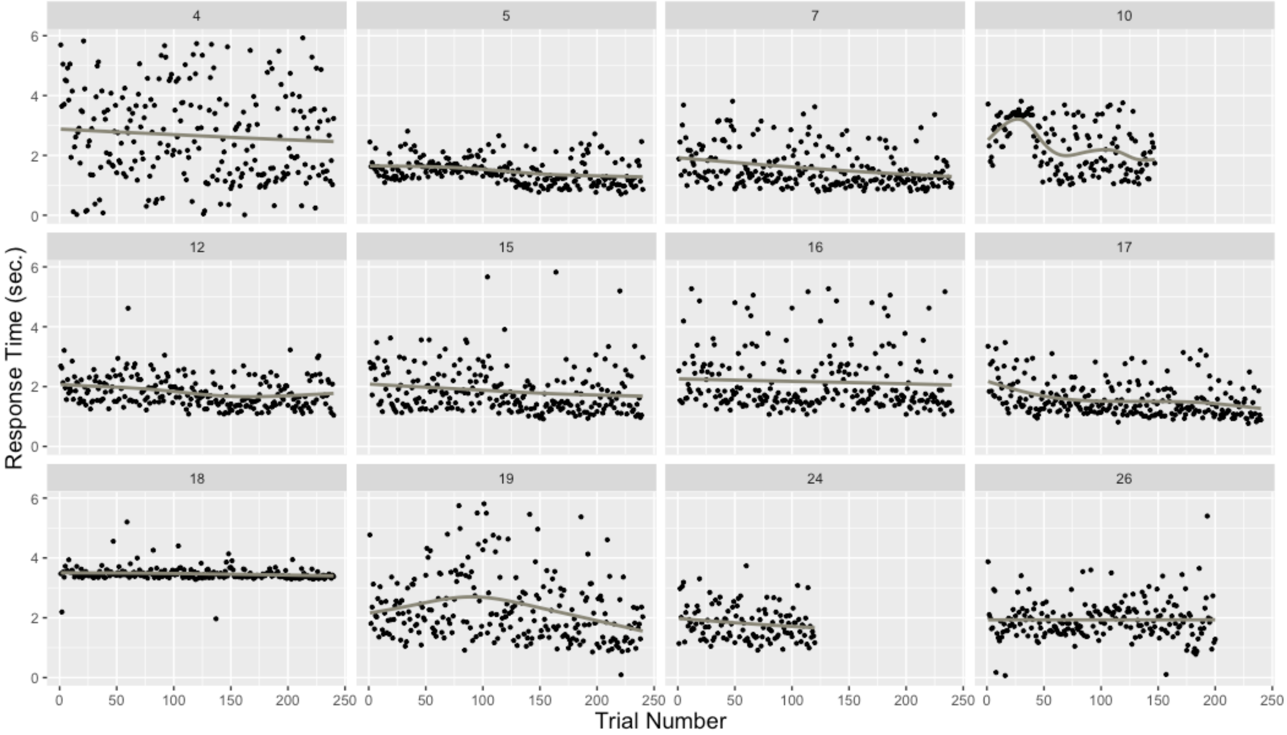
Supplemental Figure 2: Trial Number by Response Time for High Performing OA Individuals



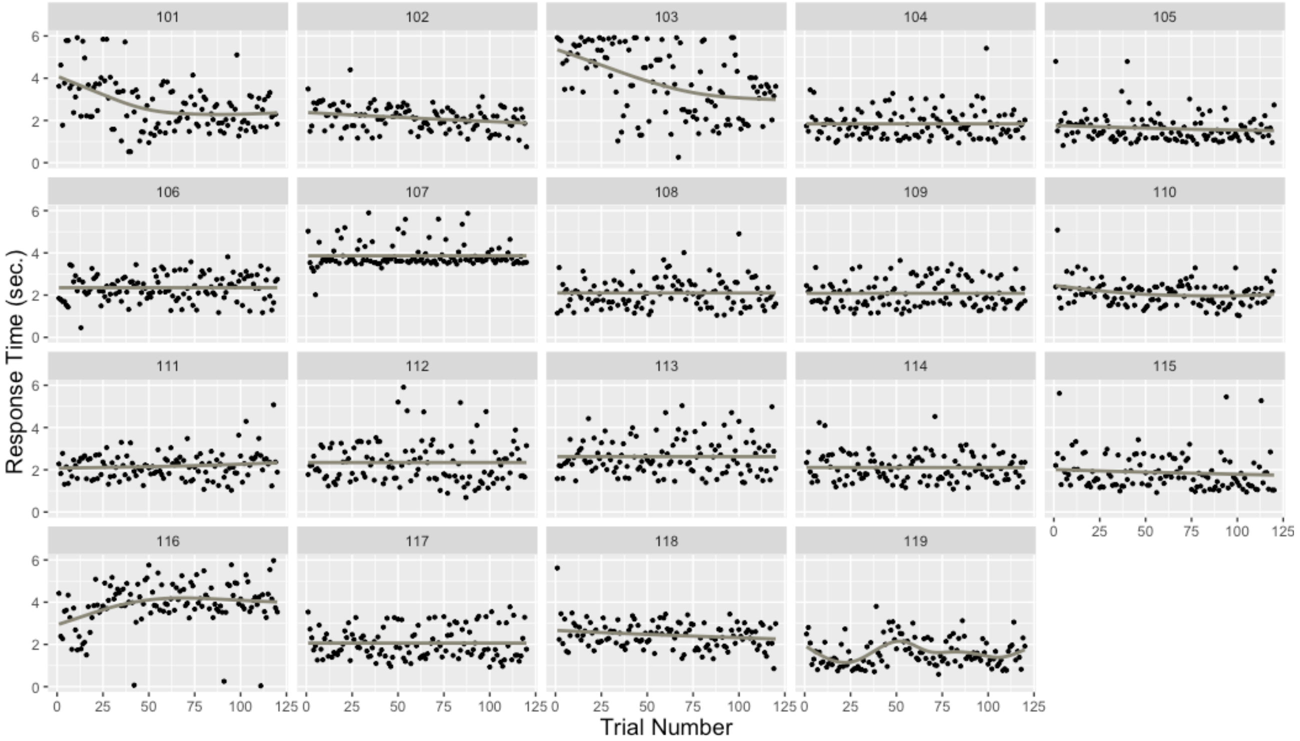
Supplemental Figure 3: Trial Number by Response Time for Low Performing YA Individuals



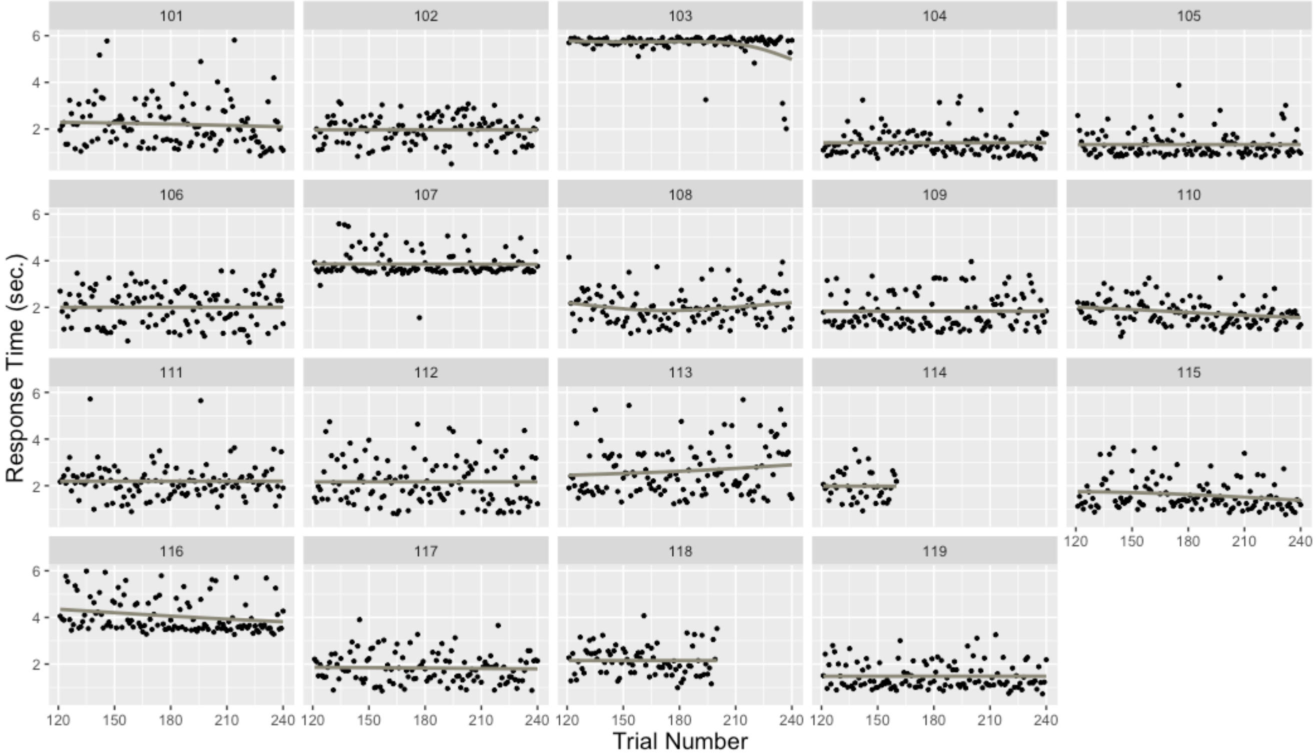
Supplemental Figure 4: Trial Number by Response Time for High Performing YA Individuals



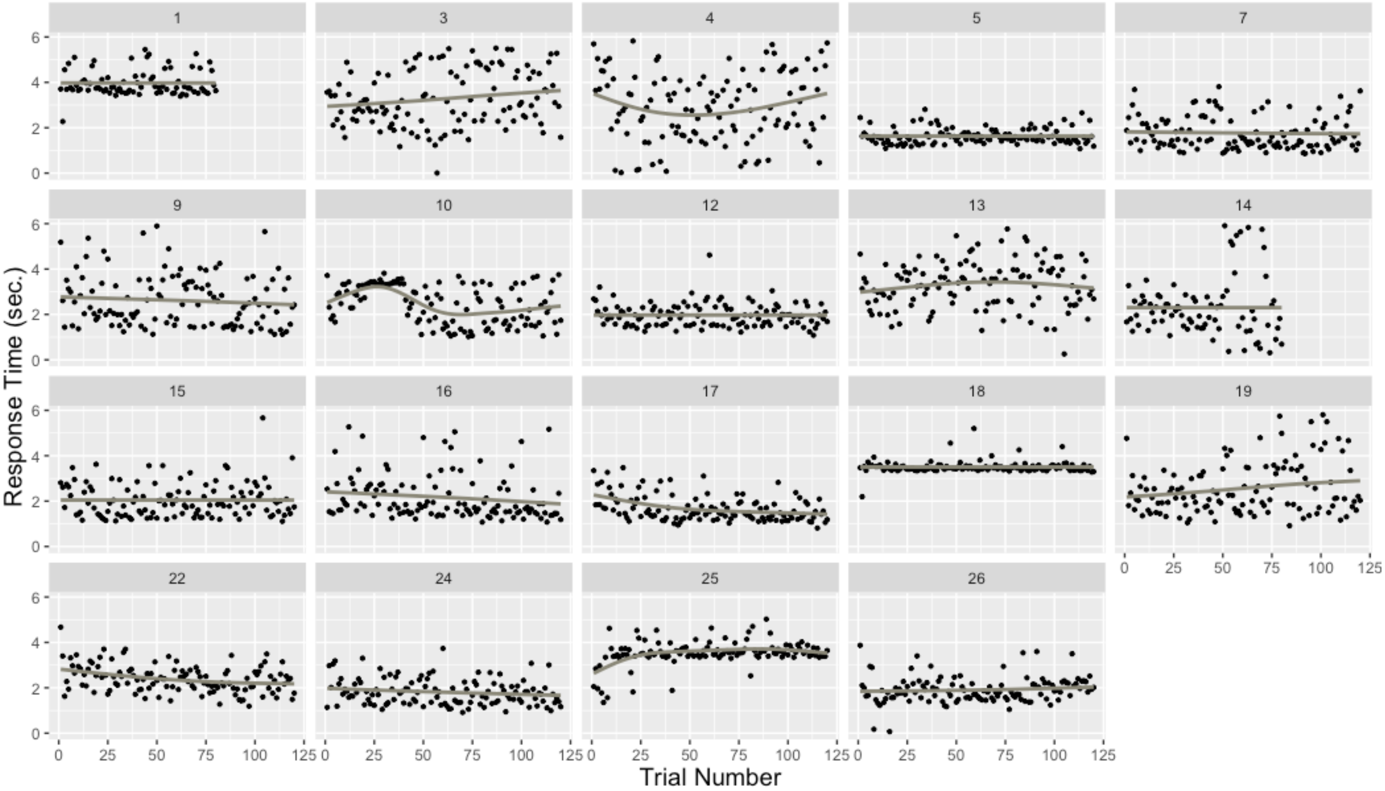
Supplemental Figure 5: Trial Number by Response Time for Early-Trial OA Individuals



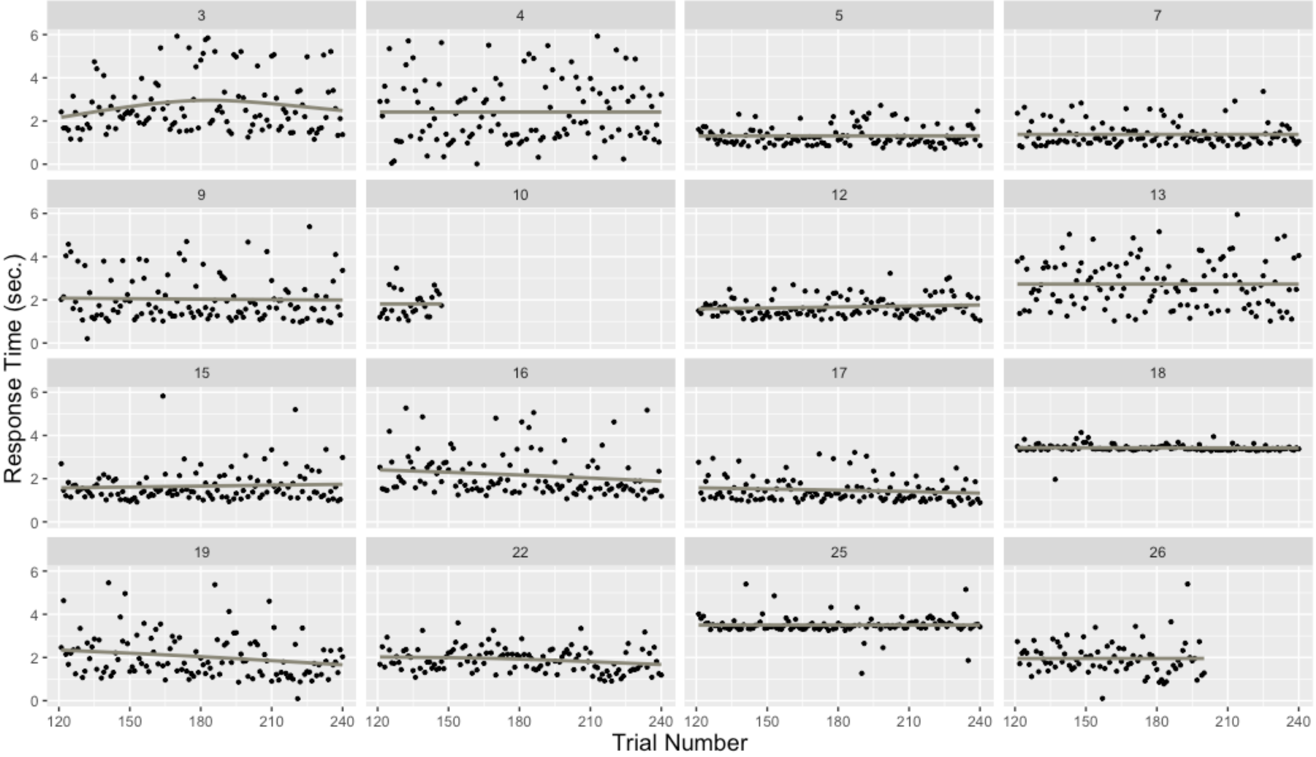
Supplemental Figure 6: Trial Number by Response Time for Late-Trial OA Individuals



Supplemental Figure 7: Trial Number by Response Time for Early-Trial YA Individuals



Supplemental Figure 8: Trial Number by Response Time for Late-Trial YA Individuals


Response Time (sec.)
Trial Number