Winter 12-15-2021

Evaluating Recognition Memory Models from an Individual Differences Perspective

Kyle Gramer Featherston
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS
Department of Psychological and Brain Sciences

Dissertation Examination Committee:
Sandra Hale, Chair
Andrew C. Butler
Ian G. Dobbins
Joel Myerson
Michael J. Strube

Evaluating Recognition Memory Models from an Individual Differences Perspective

by

Kyle G. Featherston

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2021
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor Sandra Hale for her enthusiasm and support throughout the past five years. Sandy, I owe you a great deal of gratitude for taking me in and nurturing my development. In addition, thanks to Joel Myerson and all the research assistants who have come through the Cognitive Development lab. Joel, I am extremely grateful that you were always there for me regardless of what project I needed feedback on. Your patient mentoring and meticulous editing of my writing has made me a better researcher and writer.

I would also like to thank the rest of my committee for taking the time to read, discuss, and provide feedback on this project. Beyond that, each member of the committee has been invaluable in helping to foster my learning and growth in graduate school. Ian, thanks for pushing this project to be the best it can be and teaching me so much about cognitive psychology and research. Mike, thanks for your dedication to teaching and the countless statistical, teaching, and life lessons you have taught me. Andy, thanks for letting me sit in on your class and demonstrating how cognitive research can make an impact outside of the lab.

I am grateful for the opportunity to have been a part of Washington University and the Psychological and Brain Sciences department. It has been an honor to learn from such great teachers and researchers. Most of all, I am grateful for the friends I made here at WashU. There are too many to list, but their support and their company made graduate school more fun and meaningful. Special thanks to my partner Deborah, for her support over the crazy past two years —especially during the times when stress has made me less than my usual chipper self.

To my mom, Margot, and my dad, John, I have never been able to find enough words to thank you for all that you have sacrificed for me and my siblings, Janeen and Joe. From teaching me to

read and fostering my love of books at an early age, to trying to understand what I am doing in graduate school, your support has been behind all my educational (and all other) achievements.

Kyle G. Featherston

Washington University in St. Louis

December 2021

Dedicated to my parents.

ABSTRACT OF THE DISSERTATION

Evaluating Recognition Memory Models from an Individual Differences Perspective

by

Kyle Featherston

Doctor of Philosophy in Psychological and Brain Sciences

Washington University in St. Louis, 2021

Professor Sandra Hale, Chair

Although recognition memory models have been thoroughly compared in various recognition memory paradigms, the relative reliability and validity of their parameters have not been thoroughly assessed using an individual differences approach. In two studies, I evaluated three models: the dual-process signal detection (DPSD) model, the continuous dual process (CDP) model, and the unequal variance signal detection (UVSD) model. In Study 1, participants performed a remember-know procedure that also included confidence ratings. When model parameters were estimated twice in the same individual, both key parameters from the DPSD model were reliable within an individual, whereas the CDP version of familiarity was not reliable (ICC < .40). Fitting the UVSD model also produced reliable parameters, although the variance parameter was only moderately so. In Study 2, participants performed tests of fluid intelligence, processing speed, and recall along with the same recognition procedure as in Study 1. Structural equation modeling comparing the models' ability to predict cognitive variables suggested that the parameters accounted for an equal proportion of variability in gF. However, the DPSD was the lone model with two parameters that predicted variance in gF Assessing the reliability and construct validity of the models' parameters within an individual differences' framework provided a novel test of these models. Together, the results from these studies suggest that the DPSD is the most reliable model and exhibits convergent validity with other cognitive

constructs, but that there is room for further assessment of the UVSD and the DPSD using an individual differences approach.

.

Chapter 1: Introduction

The ability to remember information and experiences from our past is one of the most important abilities humans possess. Memories of events from the past are termed episodic long-term memory (LTM). Anecdotally, these memories seem to take different forms. Our memories for some events are detailed with the time, place, sight, and smells of an occasion. Other memories bring the sensation, sometimes quite strong to the point of near certainty, that we have seen something before, but with no memory of when or where. The experience of these two types of memories has led to considerable research of potential dual processes influencing memory tasks in the laboratory and the development of dual-process models of LTM. In particular, researchers have argued that recall, source, and associative memory require a deeper memory process, termed recollection, whereas item recognition tasks can rely on both recollection and the other type of memory, termed familiarity (Jacoby, 1991; Yonelinas, 1994). This view has been termed dual-process theory.

Researchers have attempted to dissociate these processes using different task types and procedures, but the debate regarding the nature of recollection and familiarity, as well as whether both processes are needed to explain recognition memory, continues (see Rotello, 2017; Yonelinas & Parks, 2007, for reviews). In comparing models of recognition memory, differences in performance between individuals have often been ignored or treated as random variance, with some exceptions (cf. Cohen et al., 2008; Dobbins et al., 2000; Jang et al., 2009; Rotello & Macmillan, 2006; Smith & Duncan, 2004). However, determining whether individuals' performance on a memory test is related to their performance on other cognitive tasks (e.g., are those who are better than average on test A also better than average on task B) may help further our understanding of the processes involved (Underwood, 1975).

Dual-process theory has developed into one of the most cited theories of recognition memory. Dual-process theory is not, however, one uniform theory, as the exact definition of recollection and familiarity differs across theorists, but many similar theories that all share the idea that there are two processes affecting recognition judgments. For decades, various researchers have argued that there is more than one basis for judgments on recognition memory tasks (Mandler, 1980; Jacoby & Dallas, 1981; Tulving, 1985). For example, Tulving (1985) argued that the two processes influencing recognition judgments are semantic memory and recollection. Jacoby (1991) argued that one type of process (familiarity) is automatic whereas the other type (recollection) is controlled, and he developed a process-dissociation procedure for dissociating these automatic and controlled memory processes. Building on Jacoby's model, Yonelinas (1994) proposed a dual-process signal detection model (DPSD) in which familiarity is a signal detection process, whereas recollection is a threshold whereby qualitative information about a studied event (e.g., where the item was studied) is either retrieved or it is not retrieved. Recollection is measured as the probability that subjects correctly recollect some aspect of the study event. If recollection fails, then recognition must be based on a familiarity assessment process that can be described by an equal variance signal detection model and measured as d'.

Notably, Wixted and Mickes (2010) rejected the threshold interpretation of recollection and instead modeled both recollection and familiarity as two continuous processes that are summed together. Wixted and Mickes argued that when making a decision about a recognition test item, a participant takes both recollection and familiarity into account. This differentiates their model from other dual-process models that argue for a threshold recollection processes.

Although dual-process theories have received support, alternative models of recognition performance that only require a single memory-strength process have been demonstrated to be equally, or arguably more, effective at fitting recognition memory data (e.g., Dunn, 2008; Slotnick & Dodson, 2005). These alternative models and their ability to fit recognition memory data are discussed below, and a brief overview of some of the contributions that theories have made to applied areas of research such as cognitive neuroscience and aging is presented.

1.1 Models of Recognition Memory

The dual process signal detection theory (DPSD) proposed by Yonelinas (1994) and colleagues has guided much of the research in aging and neuroimaging, either explicitly or implicitly through the assumptions and tasks chosen by researchers (Wixted & Mickes, 2010). There are three main types of estimation procedures used to estimate recollection and familiarity based on the DPSD. First, the remember-know (R/K) procedure developed by Tulving (1985) asks participants to not only identify whether an item is old but qualify their response as due to "remembering" or "knowing" and sometimes "guessing". Tulving (1985) proposed that autonoetic awareness, characterized by an awareness of an event as a part of one's own past experience, was linked to the episodic memory system. Because "remembering" is associated with a conscious awareness, participants can accurately report if memory decisions are based on autonoetic awareness, or "remembered". According to this account, "know" recognitions lack autonoetic consciousness and are instead associated with noetic consciousness, characteristic of semantic memory. Later researchers have proposed that given instructions that emphasize "remembering" requires memory of the context that the stimulus was seen, "remember" responses qualify as recollection, while "know" responses indicate a lack of explicit recollection and hence must be based on familiarity, according to the DPSD (Yonelinas & Jacoby, 1995).

Second, Yonelinas (1994), developed a procedure to derive DPSD estimates of recollection and familiarity based on a receiver operating characteristic (ROC) curve from recognition confidence-ratings. In a recognition task with confidence ratings participants are asked for their level of confidence, typically with a 6-point scale, regarding their old-new decisions. These ratings and decisions then are used to create a receiver operating characteristic (ROC) curve from which estimates of recollection and familiarity can be derived. The DPSD assumes that recollection leads to maximum confidence in recognition tasks. Thus, maximum confidence hits are associated with recollection and the probability of recollection can be estimated as the intercept of the ROC. Familiarity is indexed by a signal detection process for the remaining items that are not recollected and is measured in *d´* units.

Third, the process-dissociation (PD; also known as the opposition procedure) procedure involves two recognition tests: an inclusion test and an exclusion test. In the inclusion test, a participant can rely on both recollection and familiarity. In the exclusion test, in contrast, only recollection can be relied on, whereas relying on familiarity will lead to false alarms. In the most common version of this procedure, two lists are presented, and the inclusion test instructs participants to respond “old” to any item that was presented on either list. In the exclusion test, participants respond “old’ only to items that were on a particular list, generally a spoken list that does not match the domain of the visual recognition test (Jacoby, 1991). If one assumes that recollection is based on intentionally controlled responding, whereas familiarity cannot be controlled in the same way, it is assumed the exclusion test cannot be based on familiarity. The exclusion and inclusion scores can then used to calculate estimates of recollection and familiarity. Yonelinas (2001) demonstrated that the three procedures outlined (R/K, ROC, &

PD) produced similar estimates of recollection and familiarity and converging conclusions about the effects of experimental manipulations, forming a basis for a coherent dual-process theory.

Many studies evaluating the role of recollection and familiarity have been based implicitly on a particular definition of dual processes; namely, that these three procedures accurately capture recollection and familiarity. For example, studies have used the R/K procedure to evaluate different brain regions (Kim, 2010) and to evaluate aging and amnesia (Lombardi et al., 2018). The results from these studies were interpreted as evidence for separation of brain regions and aging rates recollection and familiarity, which relies on the assumption that the R/K procedure accurately captures these two processes. However, there are other models of recognition memory that have received support in the literature, and these models have suggested both different parameters from traditional recognition with confidence tasks, as well as rejected the idea that the R/K procedure and other procedures described above accurately separate recollection and familiarity.

There have been two main classes of arguments against the DPSD model. The first is that other models may fit ROCs based on recognition memory with confidence and their z-scored transformations (zROCs) better than the DPSD (e.g., Glanzer et al., 1999; Heathcote, 2003; Heathcote et al., 2006; Wixted, 2007). The other argument is that there is evidence that the model's definition of recollection and familiarity either overstate the number of processes needed (e.g., Berry et al., 2008; Dunn, 2008;) or oversimplify the true nature of the dual-processes (e.g., Wixted & Mickes, 2010). Yonelinas (1994) first argued that ROC data, particularly the asymmetrical nature of ROC curves that are consistently found in recognition memory experiments can be explained by a probabilistic recollection process and a signal detection familiarity process. Although the DPSD model consistently provides a better fit than a traditional single-parameter single detection-model (equal variance signal detection model, Yonelinas, 1994), an alternative model with a single memory strength variable can also fit the data well (e.g., Heathcote, 2003). This unequal variance signal detection model (UVSD) has received support in the recognition memory literature based on its ability to fit ROCs and zROCs from various paradigms (e.g., Wixted, 2007). The UVSD model interprets confidence data from recognition tasks as a function of a single memory strength variable, as opposed to unique contributions from recollection and familiarity, and assumes that the signal strength of targets not only has a different mean from lures, but the variance for targets and lures also differs.

The UVSD model is defined by the unequal variance of targets and lures, as illustrated in Figure 1.1. Although the UVSD proposes a single memory strength process, the inclusion of a separate variance parameter that indexes the ratio of the variance of the target distribution relative to the lure distribution means UVSD is not more parsimonious than DPSD; the two models have the same number of free parameters for explaining ROC data.

Several researchers have pitted the DPSD model versus the UVSD model in fitting of ROC and zROC curves from various paradigms and have suggested that the UVSD model provides a better fit (Glanzer et al., 1999; Heathcote, 2003; Heathcote et al., 2006; Wixted, 2007). In particular, the UVSD predicts that the zROC ought to be linear. In the UVSD model, the value of the z-ROC slope represents the ratio of the variance of new items to lures. The DPSD suggests a slight U-shaped zROC, particularly in situations when recollection is heavily relied on, because recollection is a threshold process. If recollection is exclusively relied on, the hit rate and false alarm rate increase proportionally as the response criterion is relaxed, producing a linear ROC and a corresponding U-shaped zROC. Because the UVSD model only includes a signal detection component and not a threshold component, it does not predict a U-shaped zROC. Several studies have found linear zROCs, even in situations where recollection should play a large role, such as in source recognition (Heathcote, 2003; Slotnick & Dodson, 2005). However, DPSD theorists have argued that the evidence from ROCs and zROCs show the DPSD model provides a good fit as well and, in some instances, a better fit, and that a close to linear zROC is consistent with the DPSD model (Yonelinas & Parks, 2007).

In regard of studies that demonstrated a better fit to the data for the UVSD, proponents of the DPSD theory have argued that the overall model fit is quite good for both models, and therefore does not provide conclusive evidence for one model over the other (Yonelinas & Jacoby, 2012). Indeed, comparison of models' ability to fit a variety of recognition memory data post-hoc primarily assesses the flexibility of the models, even if penalties for number of parameters are applied. However, these comparisons do not provide much information on the usefulness of the models' parameters to understand psychological processes or broader cognition.

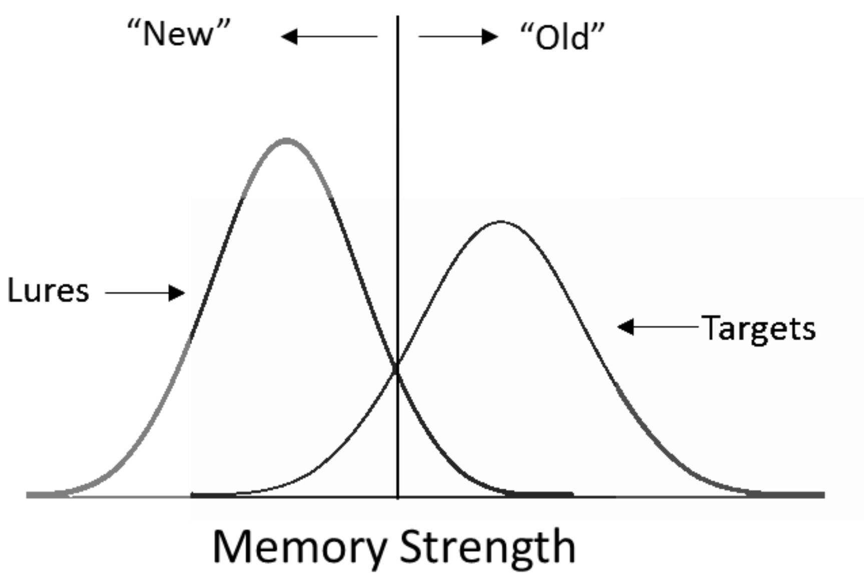


Figure 1.1. An illustration of the unequal variance signal detection (UVSD) model modified from Wixted (2007).

In addition to explaining recognition memory ROCs, there is some evidence that some of the other findings in recognition memory research can be explained by a single memory-strength variable. Dunn (2008) used state-trace analysis, which is designed to evaluate the number and organization of variables that mediate the effect of independent variables on two or more dependent variables. Based on data from 38 studies that used the R/K procedure he concluded that only one processes was needed. Specifically, the data from the studies suggested 'Remember' and 'Know' judgments were affected similarly by various experimental factors, so that when the remember rate increased the know rate also increased and vice-versa These results matched the predictions of the UVSD model because it lacked evidence for two processes that are affected separately by different experimental manipulations. However, this study did not entirely rule out that two processes could be responsible for the data. In addition, by using overall memory performance across individuals in various studies, as with studies that have evaluated ROC fits across participants, the method did not evaluate the effects of manipulations on individual participants' remember or know rate, or the corresponding recollection and familiarity estimates.

The exact definitions of recollection and familiarity proposed by the DPSD model have also been criticized. Specifically, the characterization of recollection as an "all-or-none" retrieval process and the reliance on recollection for all relational memory judgments (e.g., source and associative memory) have been questioned. Findings have suggested that recollection is not all-or-none, such that confidence in recollection judgments can be graded (Ingram et al., 2012; Rotello et al., 2005). Yonelinas and Parks (2007) have clarified that the DPSD model does not necessitate that recollection is all-or-none in the sense that everything about an event will be recalled or nothing; clearly some contextual details of an event can be recalled and not others. Instead, the hypothesis is that there is a threshold, such that if a certain level of contextual detail is remembered one experiences a *recollection* of the event itself and responds with high confidence. Recollection being associated with different levels of confidence is therefore not necessarily viewed as contradicting the DPSD model. However, the DPSD model does argue that because recollection is a threshold process, measuring it as a yes/no probability is sufficient.

Similarly, relational judgments, such as memory for context or source, have been shown to be above chance for memories that the DPSD would characterize as familiarity-based, such as "know" ratings (Wixted & Mickes, 2010). A strict interpretation of the DPSD suggests that relational judgments should be based entirely on recollection. However, a degree of relational memory associated with familiarity judgments does not rule out the major assumptions of the DPSD, as these can be thought of as relational memories that do not cross the threshold of recollection.

In the UVSD model, recollection and familiarity processes are not explicitly modeled. However, due to the contribution of dual-process models in explaining a variety of recognition memory paradigms, some advocates of the UVSD argue that the UVSD can be successfully modified to include contributions of recollection and familiarity. Researchers have extended the UVSD model by arguing that recollection and familiarity, rather than independently contributing to recognition decisions, additively contribute to the overall memory strength variable (Wixted & Stretch, 2004; Rotello et al., 2005). Wixted and Mickes (2010) formalized this theory, termed the continuous dual-process model (CDP). The CDP proposes that when making a decision about a recognition test item, a participant takes both recollection and familiarity into account. This renders it impossible to separate contributions of recollection and familiarity from each other in a standard recognition task with confidence; in such a task, the UVSD and CDP models will be identical. Additionally, the theory assumes that participants can attribute decisions primarily due to recollection or familiarity in a traditional R/K procedure, but because recollection and familiarity both contribute to a given recognition decision, the decisions are assumed to not be process-pure and thus the standard procedure will not effectively separate the two processes. In addition, because both recollection and familiarity vary in strength, confidence ratings are needed to accurately calculate each process. An illustration of how both the DPSD and CDP suggest remember-know and confidence judgments are made is shown in Figure 1.2.

CDP researchers have utilized a task in which continuous recollection can be separated from familiarity. This task is a combination of an old/new recognition memory test with the R/K procedure (Rotello et al., 2005). Participants are asked to rate their confidence that a test item is old, and if the item is judged to be old, they also make a Remember or Know judgment. Using this task, Wixted and Mickes (2010) found evidence that recollection judgments based on different confidence levels do vary in item accuracy, consistent with results from Rotello et al. (2005). In addition, although high confidence 'Remember' and 'Know' judgments did not differ in item accuracy, 'Remember' judgments were associated with higher source accuracy. This was evidence that there was a distinction between recollection and familiarity, unlike what a single process model would suggest. However, source memory judgments were above chance for high confidence familiarity decisions, suggesting that recollection and familiarity are not entirely separable (or that source memory judgments can be based partially on familiarity).

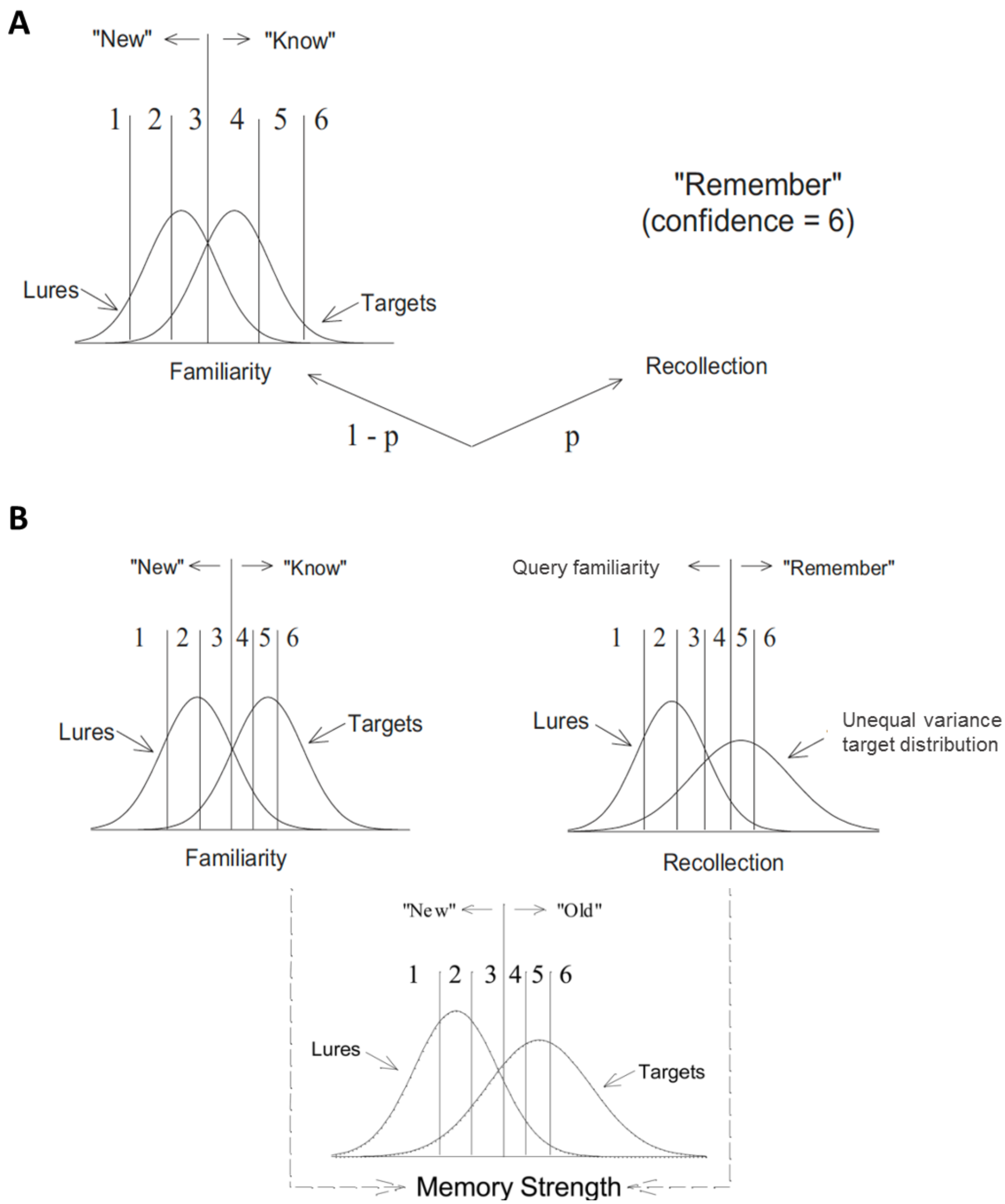


Figure 1.2 Modified from Wixted and Mickes (2010). A. An illustration of the dual-process signal detection (DPSD) model for making confidence and remember-know judgments. The probability that recollection occurs for a test item is p. B. An illustration of the continuous dual process model (CDP) for making confidence and remember-know judgments. In the illustrated model, recollection and familiarity are queried separately to make remember and know judgments. To make an Old/New decision, recollection and familiarity signals are aggregated to form a continuous memory strength variable.

A subsequent study that asked participants to make simultaneous confidence and familiar versus remember decisions demonstrated that when medium confidence recollection decisions were compared to high confidence familiarity decisions, they were more accurate with regard to source accuracy but equally to less accurate on item accuracy, both when a confidence scale ranging from 1-20 and from 1-6 was used (Ingram et al., 2012). This study did not formally fit the CDP model but the findings from this and other studies using a combination of R/K and confidence tasks (Wixted & Mickes, 2010) are consistent with the CDP model's predictions. Namely, the data suggest that high confidence item decisions can be made on the basis of familiarity, that recollection can vary with confidence, and that recollection and familiarity are different, where recollection is categorized as greater memory for the source or context in which an item was studied. As discussed above, the findings of varying confidence recollections and above-chance source judgments based on familiarity do not invalidate the DPSD (Yonelinas, 2007). However, they do suggest that calculations of recollection and familiarity based on the DPSD may be oversimplified, and in particular that factoring confidence of remember judgments may better capture recollection.

In summary, a large number of studies have attempted to determine the model that better fits ROCs from recognition data. Although some argue that the evidence favors the UVSD model (e.g., Heathcote et al., 2006), others disagree and argue that the DPSD is equal to better (Yonelinas & Jacoby, 2012). This has led to the development of a third model that attempts to combine the two models (Wixted & Mickes, 2010). As Roberts and Pashler (2000) argue, however, simply evaluating whether a model can fit data is not a good way for establishing a useful theory that drives progress. By evaluating the relative fits of different models post-hoc, researchers are not testing specific competing predictions of the models. In addition, the emphasis on model fitting benefits more flexible models and does not evaluate whether there are even plausible results that the theory cannot fit. Research on recognition memory models has arguably over-emphasized fitting data from specific tasks to the detriment of a broader understanding of memory (Hintzman, 2011). If each model fits recognition data relatively well, determining which is the most compelling model might not be best answered by which is more accurate in describing recognition data, but rather which is more useful in explaining memory and behavior more broadly. Given the ongoing debate as to which recognition memory model best fits recognition memory data, it is important to remember that these models are highly simplified descriptions of complex cognitive processes that underlie recognition memory decisions. Creating a model that perfectly fits the data from all individuals in all conditions is not possible and perhaps not desirable because such a model would require so many parameters as to not be useful. Evaluating these models' usefulness in understanding individual differences could provide a benefit in our broader understanding of memory by evaluating psychometric properties of parameters such as reliability and construct validity.

1.2 Individual Differences in Recollection and Familiarity

Underwood (1975) argued that individual differences are necessary for theory construction, as theories ought to make specific predictions about the correlations between one or more tasks or group of tasks, and this logic applies to recognition memory models. This is particularly true because dual-process theory's applications in areas such as neuroimaging and aging either suggest or are based on the explicit operationalization of recollection and familiarity as reliably different between individuals. For decades, psychologists, have argued that assessment of individual differences is an important contributor to constructing and assessing theories (Eysenck, 1984; Underwood, 1975). As statistical and methodological techniques have continued to evolve, evaluating individual differences continues to be important to evaluate and compare theories of cognition (Goodhew & Edwards, 2019; Hedge et al., 2018; Rouder & Haaf, 2019). Here, previous work exhibiting the usefulness of dual-process models in cognitive neuroscience and cognitive aging research, and how this work relies on assumptions about reliable individual differences across individuals and occasions, is overviewed.

Presumably if different brain regions are implicated in recollection and familiarity, then individuals will differ in the function of these regions. A growing body of neuroimaging research has identified different regions associated with the purported dual processes. For example, functional MRI studies have identified different areas of the medial temporal lobes (MTL) that are associated with the two processes at retrieval, with the hippocampus in particular being associated with recollection (Daselaar et al., 2006). Dissociations have not been found solely in the MTL, which has long been associated with storage and retrieval of LTM, but also a range of brain regions. In particular, the brain regions associated with recollection have received considerable interest, and recent research has converged on a "core recollection network" including the left angular gyrus, medial prefrontal cortex (mPFC), hippocampus, middle temporal gyrus, and posterior cingulate cortex (Kim, 2010; King et al., 2018). These studies identify specific brain regions that appear to be activated during recollection in different tasks and studies, and if the same brain regions are activated for recollection across occasions, then it should be possible to measure recollection reliably in an individual across occasions.

In contrast, perirhinal activation has been found consistently for familiarity- based judgments (e.g., Daselaar et al., 2006, Gonsalves et al., 2006; Uncapher & Rugg, 2005). Importantly, the qualitative differences between the brain regions activated in recognition memory were determined based on predictions of dual-process models. Namely, the findings have shown that different regions are activated by "remember" versus "know" decisions, by source and associative memory tasks versus item memory tasks, and by high confidence versus lower confidence decisions (see Eichenbaum et al., 2007, for a review). These results underscore the utility of the dual-process model in contributing to findings in neurocognitive research by making specific predictions about distinctions in brain region functions. Further, the nature of the findings emphasizes that other models of recognition memory ought to not only be able to explain these results, but ideally also make unique predictions of their own.

Although it is possible that for healthy younger adults, the difference in function of these regions is negligible and does not result in reliable individual differences, in at least one study, functional connectivity of the core recollection network predicted individual differences in memory accuracy in young adults (King et al., 2017). This suggests that there are meaningful individual differences in the function of these regions as they relate to performance even in younger adults.

Research in cognitive aging has also been influenced by dual-process theories of recognition memory. This research, along with similar research in amnesia and dementia patients (Koen & Yonelinas, 2014), suggests that certain populations of individuals differ in their ability to utilize recollection and familiarity. Jacoby and colleagues have proposed a hypothesis that recollection declines with age, but not familiarity. In one study, Jennings and Jacoby (1997) demonstrated that older adults performed worse on a recognition test only when they needed to differentiate words that were shown at study versus those shown during previous testing (i.e., when needing to rely on recollection). Similarly, recall has long been suggested to decline with age at a greater rate than recognition (e.g., Craik & McDowd, 1987).

A meta-analysis of studies that included both a recognition and recall measure for both younger and older adults found that recall showed a reliably greater deficit in older adults than recognition (Rhodes et al., 2019). However, the effect size for recognition was greater than zero (*d* = 0.544 compared to 0.891 for recall) indicating that there is decline in both types of memory. These results are consistent with a decline in recollection but not familiarity with age, as recollection operates in both test types, accounting for the greater than zero decline in each, but familiarity plays a larger role in recognition, accounting for the smaller decline. Other studies have directly estimated recollection and familiarity and found larger deficits in recollection than familiarity based on procedures such as the process dissociation procedure (Jennings & Jacoby, 1997), the R/K procedure (McCabe et al., 2009) or a combination of procedures (Koen & Yonelinas, 2016).

The argument that a wide range of older adult deficits in memory tasks can be explained by deficits in recollection, but intact familiarity, relies on the assumption that recollection and familiarity are not just processes that explain recognition performance, but variables that differentiate performance on multiple tasks. If recollection and familiarity are not reliable within an individual across different measures, then describing recollection as "declining" in older adulthood does not make sense. This argument relies on the idea that recollection is indeed an ability within individual, that presumably can be measured in a consistent way in both younger and older adult populations. The recollection-deficit hypothesis' usefulness as a cognitive aging theory is dependent on the ability to measure recollection (and familiarity) across time and measures. Similarly, if one rejects dual-process models, a useful alternative model should offer an explanation as to why performance on certain LTM tasks declines more rapidly with age, and how a specific age deficit can be measured reliably.

However, surprisingly few studies have investigated individual differences in recollection and familiarity, whether based on the DPSD or the CDP. Similarly, estimates of parameters based on the UVSD model, namely d' and variance of the target distribution, have not been compared directly to dual-process theories with regard to reliability or predictive validity in individuals. Given the wide applications of recollection and familiarity in areas such as neurocognition and aging research, establishing that they are variables that can be reliably measured for a given individual is an important step that has not received the attention in the literature that it deserves. A goal of this dissertation is to rectify this shortcoming and provide a more comprehensive evaluation of recollection and familiarity parameters, as well as UVSD parameters, as individual differences variables by fitting different models to data from the same participants.

A small number of studies have investigated individual differences in recollection and familiarity. Wang and Yonelinas (2012) attempted to evaluate whether individual differences in performance on implicit memory tasks separately predicted estimates of familiarity and recollection. Priming scores from a free association task were strongly correlated with familiarity scores from a ROC procedure based on the DPSD model (r = .48) and a R/K procedure (r = .56) but were not significantly correlated with recollection. In an additional experiment, a similar procedure was used, but the free association task was replaced by an associative cued-recall task where participants were asked to try to specifically recall words that had already been studied when presented with a categorically related priming word. The inverse pattern was observed; the proportion of previously seen words recalled was significantly related to recollection, but not related to familiarity (Wang & Yonelinas, 2012). This study suggests that familiarity and recollection are reliable individual differences for a particular list of words, as those who demonstrated higher scores on familiarity in a recognition task tended to retrieve those same words in a separate implicit memory task and those who exhibited greater recollection recalled more of those words when given explicit instructions to do so. However, it did not demonstrate that recollection and familiarity scores were reliable within an individual across multiple lists of items, as would be expected if recollection and familiarity are cognitive abilities. In addition, parameters from competing models such as the UVSD were not tested, so whether similar dissociations could be formed such that the variance and discriminability parameter may relate differently to recall and implicit memory.

A common, but relatively unused in recognition model comparison, approach for establishing individual difference constructs across multiple measures comes from latent variable analysis, such as confirmatory factor analysis (CFA), a form of structural equation modeling (SEM), because latent variables represent the shared variance among multiple measures and account for measurement error. Unfortunately, most of the extant latent variable analyses of recollection and familiarity have suffered from lack of power. For example, a study recruited hypoxic patients to test the hypothesis that damage to the hippocampus in hypoxia would affect recollection but not familiarity (Quamme et al., 2004). Although the use of a unique population limited their sample to 56 participants, CFA results suggested that the best-fitting model separated recollection and familiarity, defined by having both recall and recognition tasks load on recollection and only the leftover variance of recognition task load on familiarity (Quamme et al., 2004). Another study by Koen and Yonelinas (2016) had participants perform three different recognition procedures: R/K, ROC, and PD. The results indicated a moderate correlation between recollection estimates from the three tasks (*r*s = .48, .55, 68). The same was true for two of the familiarity correlations but familiarity estimates from the ROC and PD procedures were not significantly correlated. Additionally, a CFA was performed with hits from the three tasks loading on both factors, and false alarms loading only on familiarity, with an additional recall task also loading on recollection. This model indicated a better fit than a single factor model, but these results were based on only 39 participants. Therefore, these findings should be considered preliminary since they fall well short of recommended sample sizes for multi-factor models (e.g.. Wolf et al., 2013).

The most comprehensive individual differences study of recollection and familiarity to date had young participants perform recall, item recognition, and source recognition tasks (Unsworth & Brewer, 2009). They tested a dual-process CFA model in which a recollection factor included performance on all tasks, and familiarity included only the residual variance from item recognition. This model had an excellent fit, fitting the data better than either a single factor model or a model that separated recognition and recall tasks (Unsworth & Brewer, 2009). This study provided strong evidence that recollection and familiarity are distinct abilities in which individuals differ and suggests that these abilities explain variance in performance on a wide variety of LTM tasks. However, the study did not specifically offer evidence of the reliability or validity of the parameter estimates of familiarity and recollection from recognition memory procedures (i.e., R/K, ROC, & PD). It also did not compare the CDP model with the DPSD, because both models posit that source memory and recall rely primarily on recollection.

None of the above studies compared the DPSD to the CDP or the UVSD model. Individual differences researchers have tested single process models by single-factor LTM models consisting of shared variance across many different tasks (Unsworth, 2019) but not the explicit parameters of the UVSD model. One study of patients with hippocampal damage calculated parameters from both the UVSD and the DPSD model and found that hippocampal damage was associated with a decrease in both the memory strength and variance of targets parameter compared to controls (Dede et al., 2013). The results suggested that Variance of the target distribution is correlated with recollection, however from a small sample size (5 participants with hippocampal damage and 14 control participants). Outside of this study, proponents of the UVSD model have not offered much explanation of what the Variance parameter represents in terms of cognition or individual differences.

A few studies have compared model fits for individuals' ROC data and argued that the UVSD is a better fit for individual data (Heathcote, 2003; Healy et al., 2005) although the results of other studies using the same method suggest that the DSPD provides better individual fits (cf. Howard et al., 2006). Moreover, the preceding studies still suffer from the same problems that other studies emphasizing fits to single recognition memory tests do. They say nothing about the reliability of the estimates within an individual nor the usefulness of the estimates in explaining other areas of cognition. Although Smith and Duncan (2004) did attempt to evaluate the reliability of the UVSD compared to the DPSD in two different recognition memory paradigms (two alternative forced-choice and confidence ratings), as Yonelinas and Parks (2007) subsequently pointed out, that study was not a valid test of the DPSD because participants were explicitly instructed to use all points on the confidence scale equally. The DPSD posits that high confidence old responses are more common than lower confidence responses, particularly when performance is high, since recollection leads to a high confidence response. Thus, if an individual recollected more than 1/3 of the old items, for example, they would not have been able to respond with high confidence for all their recollections, leading to a poor estimate of their recollection. Further research on this issue is clearly needed because fitting group data can obscure patterns in individuals' data and imply misleading conclusions (e.g., Brown & Heathcote, 2003).

1.3 Individual Differences in LTM

To date, there is still not a consensus in the literature as to the structure of LTM abilities. Several studies have focused on the relation between various LTM tasks, and frequently found that there is considerable shared variance among tasks, but there is also unique variance for different task-types. For example, Unsworth (2010) found that across a number of different recall and recognition tasks, the best-fitting model separated recall and recognition. Other studies suggest that associative memory (Malmi et al., 1979) or source memory (Unsworth & Brewer, 2009) are unique constructs from either recall or recognition, although yet other studies suggest otherwise (Siedlecki, 2007).

In a recent review paper, Unsworth (2019) reanalyzed data from multiple studies and found that the best-fitting model separated task-specific factors (e.g., paired associates, source memory, recall, and recognition), although these factors shared enough variance to suggest a higher-order LTM ability. This evidence was inconclusive regarding recollection and familiarity, given that the tasks included in Unsworth's analysis all rely on recollection to some extent and thus would be expected to be related, and a model based on a dual-process interpretation was not explicitly tested. Ultimately, Unsworth (2019) concluded, "we have only scratched the surface in terms of understanding individual differences in LTM." (p. 126).

Given this state of the literature, recognition memory models can contribute to our understanding of LTM by furthering our understanding of individual differences in LTM abilities. In particular, dual-process theories' use of recollection and familiarity offers specific predictions about which LTM tasks share more variance than others, whereas a UVSD model that suggests a single ability, would suggest that all tasks are similarly related. A key question regarding the models favored by many studies of individual differences in LTM is to what degree does the fit of these models reflect method variance. Different tasks are, by definition, assessed using different methods, and as a result, the unique variance of these tasks may not necessarily be particularly meaningful theoretically. Dual-process theories may not explain all the unique variability of tasks, but the degree to which tasks rely on recollection and familiarity offers potential explanations for what may be common and what may be unique among various tasks.

In addition, the estimation of recognition memory decision models offers an opportunity to go beyond task performance and understand the processes involved in specific LTM tasks, such as recognition. Previous studies of LTM using latent variables provide insight into individual differences in performance on LTM tasks, but generally rely on performance throughout an entire task. Given the fact that multiple processes generally contribute to performance on a task (Jacoby, 1991), the insight into processes is generally limited. Some researchers have attempted to address the multiple processes involved in tasks such as recognition, by having the task load onto multiple latent factors (e.g., Quamme et al., 2004). However, to my knowledge, no study has directly estimated the parameters of decision models before using those parameters (rather than measures of overall task performance) as indicator variables.

In summary, the exact number and nature of individuals' LTM abilities is yet to be determined. In particular, although recognition memory models have been thoroughly compared in recognition memory paradigms, their parameters have rarely been examined in research on individual differences. Therefore, a major goal of this dissertation was to examine the utility of recognition memory models from an individual differences perspective, assessing the reliability of model parameters and the construct validity of the LTM abilities that they imply. This research offers the added benefit of providing novel tests of prominent recognition models, while also advancing our understanding of individual differences in LTM more generally.

Chapter 2: Aims and Hypotheses

The goal of this dissertation is to evaluate recognition memory models' reliability and validity by assessing how well the models' parameters can account for differences between individuals' performances across occasions. A motivating principle of this dissertation is the aphorism that "all models are wrong, but some are useful" (Box, 1976). This dissertation will help to determine which of the recognition memory models discussed above (i.e., the DPSD, the CDP, and the UVSD) is the most useful for evaluating individual differences in cognition. First, each of the models' parameters was evaluated for their reliability in a sample of approximately 60 individuals (Chapter 3). Next, approximately 200 participants performed the recognition test along with a battery of other cognitive measures in order to assess the relationship with each model's parameters with other cognitive constructs (e.g., fluid intelligence; Chapter 4).

2.1 Reliability of Model Parameters

A necessary precursor to establishing a measurement's validity is to establish its reliability. If the same participants perform a test multiple times within a short time frame and do not score similarly each time, then the score is not a good measurement of an underlying ability. Overall performance on recognition memory tests (e.g., the number of items correctly identified as old and new) is generally relatively reliable (Bird et al., 2003). However, the reliability of parameters from each of the three decision models at issue here has not been assessed previously. A lack of reliability for an otherwise good-fitting model could be a result of a model overfitting to a particular dataset. On the other hand, if the parameters are reliable, then that indicates that the parameters are consistently measuring something. However, reliability does not guarantee that what is being measured is a meaningful cognitive attribute of an individual. Nevertheless, assessing the parameters of each model not only provides its own evaluation of the models, it also provides reliability estimates to use in further analyses. As will be described in Chapter 4, these estimates were used in calculation of reasonable error variance terms for the parameters in Structural Equation Models.

2.2 Which model best predicts fluid intelligence (gF)?

Fluid intelligence (gF), the ability to solve novel, abstract problems, is a well-studied construct (Horn & Cattell, 1966). Measures of gF whose reliability and validity have been established have been used extensively by researchers (e.g., Raven, 2000). General cognitive ability and reasoning have been shown to predict real-word outcomes, including school performance (Deary et al., 2007), job performance (Schmidt and Hunter, 1998), health (Wrulich et al., 2014), and all-cause mortality (Aichele et al., 2013; Calvin et al., 2011). Due to the demonstrated contribution of cognitive ability, and more specifically gF, to these outcomes, gF has commonly been used in individual differences research, often as a dependent variable predicted by other constructs, such as working memory capacity (e.g., Conway et al., 2002; Engle et al., 1999). Given the relation often found between cognitive constructs and given that any test of gF generally involves some form of retrieval from LTM, it is reasonable to assume that a good measure of LTM ought to predict a reasonable proportion of the variance in gF. Of course, this relation, like most relations between psychological constructs, is likely to be bidirectional, as those with greater ability to solve fluid problems will likely also be able to use this capability to remember more information in LTM.

Indeed, previous studies have found a strong relationship between LTM and gF. Unsworth’s (2019) review found that across studies using latent variable methods, the weighted average correlation between an overall LTM factor and gF was .58. Interestingly, when assessed separately, both recall ability and recognition ability have been shown to predict gF (Unsworth, 2010). If recollection and familiarity underlie recognition ability, this finding could be due to the shared influence of recollection, or it could be that familiarity is also predictive of gF.

Unsworth and Brewer (2009) found a recollection factor based on source memory, item recognition, and item recall measures that was related to both gF and WM, while a familiarity factor based on the unique variance of item recognition tasks, was not related to these constructs. The larger role of recollection is a consistent theme in the limited research evaluating recollection and familiarity's relation to other variables, such as the finding that recollection was strongly related to children's academic achievement, whereas familiarity had only a small correlation (Blankenship et al., 2015). On the other hand, a study of memory for scientific passages found that both familiarity and recollection were related to verbal ability, whereas recollection was related to prior knowledge of a topic (Long, et al., 2008). This finding suggests that familiarity may be related to other abilities and also suggests that recollection may vary within an individual depending on their prior history with memory materials, rather than being an individual ability. As in other studies of individual differences in recollection and familiarity, none of the above studies directly compared recollection and familiarity estimates based on the DPSD model with estimates from the CDP model for their ability to predict other cognitive constructs, nor did they directly compare parameters from the UVSD model with recollection and familiarity from either of these models.

Although the results of these studies are informative in helping understand recollection and familiarity as potential individual difference variables, they do not offer any insight into a specific recognition model nor model comparison. Specifically, each of the above studies grouped specific tasks that are hypothesized to be related to recollection and familiarity and formed latent models. This only indirectly tests the hypothesis that recollection and familiarity are responsible for individual differences in recognition memory and says nothing about the preferred method of estimating these variables (i.e., the DPSD or the CDP estimates). Furthermore, because no task is process-pure (Jacoby, 1991), the approach of using overall task performance on memory tasks as indicator variables limits this approach to understanding underlying cognitive processes. Instead, latent variables in traditional CFA approaches capture abilities that may be made up of several underlying processes. However, the CFA approach can also be used to understand processes if those processes are directly estimated and then included in the CFA model, as was done here.

This dissertation examined the ability of estimates of the parameters of each of the DPSD, CDP, and UVSD models to predict a latent gF variable formed by established measures of fluid reasoning ability. The basis of the overall comparison of models was the proportion of variability in gF that they key parameters of each recognition memory model explained. Recognition memory, as a measure of overall memory ability, should be predictive of gF for several possible reasons. Those with higher gF will likely be able to better choose and implement encoding and retrieval strategies, such as forming more detailed retrieval cues in recognition memory tasks. Those with high gF also generally tend to be those with greater working memory capacity (e.g., Conway et al., 2002), which should give them an advantage in organizing and storing items in LTM during study. In the other direction, those with higher recognition memory ability will likely benefit from their greater ability to recognize, through their ability to recognize and retrieve from long-term memory patterns or elements of the solution, as well as draw upon previous experience in tests of gF. example. Regardless of the exact mechanism, there is an established association between overall recognition memory performance and gF (Unsworth, 2019). Therefore, a useful model of recognition memory ability ought to be able to predict variability in gF and the models were compared on this criterion.

2.3 Defining the parameters in each model

In addition to comparing the models, the present analyses should provide further clarification of model parameters that have previously only been vaguely defined psychologically. Although *recollection* is usually defined relatively clearly, *familiarity* is often poorly defined. Hintzman (2011) argued that, "Familiarity is routinely invoked in formal and informal explanations of memory as though it were a concept with obvious meaning, but the term appears to mean more than one thing… The field could benefit from a careful analysis of the ways in which the concept of familiarity has been used.'' (p. 259). Researchers have attempted to define familiarity, but these definitions are not always in agreement and too often familiarity is defined as "not recollection" as evidenced by remember-know instructions that instruct participants to respond 'know' to any item they recognize but do not 'remember.' Although the CDP and the DPSD models compute familiarity differently, the cognitive process itself is defined similarly in both cases, and thus both models are subject to criticism of the term's definition. Assessing the relationship of familiarity to other cognitive variables reveals which specific cognitive abilities are and are not related to familiarity in order to clarify which aspects of cognition underlie the ability to make familiarity-based memory decisions. These relationships have only been assessed in a few studies (Blankenship et al., 2015; Unsworth & Brewer, 2009) and this dissertation can offer needed clarification of the nature of familiarity.

In terms of specific predictions of the analysis of gF, the recollection parameters from the DPSD and the CDP model were hypothesized to predict a significant amount of variance in gF, and the relative amount of variance explained by each was compared. Recollection has been defined as a controlled process that is dependent on capacity limitations (Jacoby, 1991), which suggests that it is susceptible to an individual's use of resources and strategies, as are tests of gF. Additionally, it has been postulated that retrieval processes in recognition are related to organizational structure of memory (Mandler, 1980). Presumably those with greater fluid reasoning will be more adept at organizing study items into memory. In addition, an ability to recollect previous experiences would increase the ability of participants to use their previous experience (both from previous trials and form extra-experimental settings) to identify the strategies and rules that have worked in solving the type of problems that occur in tests of gF. Thus, a relationship between recollection and gF is expected, as those with higher gF would be expected to score higher on recollection and vice versa.

Based on previous null relationships between a familiarity factor and gF (Unsworth & Brewer, 2009), it was hypothesized that the familiarity parameters from each model would not explain additional variance beyond recollection ability. However, this association has not received much previous attention. Although the memory process of familiarity is not necessarily well-defined, those who make more accurate decisions in the absence of recollection, will score higher in familiarity. Thus, decision making ability in noisy environments may be related to gF, and familiarity was allowed to predict gF in the current analyses.

The d' parameter is an index of the distance between the target and lure distributions and thus for an individual participant represents *general* memory strength across items. In some descriptions of this model consistent with the CDP, the d' is a summation of recollection and familiarity (Wixted, 2007). Regardless of whether it represents the summation of multiple processes or a single process, it represents an index of overall recognition ability, and as discussed above, because general recognition ability is linked with gF, d' ought to be a predictor of gF as well. The UVSD model posits that a *Variance-of-targets* parameter is needed to fit recognition data. Although this variance parameter enables the model to fit ROC data well, what cognitive process it represents is unclear. Similarly, although the CDP does not necessitate its inclusion, the CDP model previously fit in the literature includes a variance of the recollection distribution parameter (Wixted & Mickes, 2010). The variance parameter is not defined beyond the assumption that there is greater variability in the encoding and retrieval of targets than lures. One hypothesis is that the encoding of a particular item during study induces variability, such as the amount of attention paid or personal connection to a particular stimulus, that leads to the strength of each item varying from trial-to-trial and from participant-to-participant (Jang et al., 2012, Wixted, 2007). However, this particular hypothesis of the Variance-of-targets parameter has not held up well to scrutinization, as manipulations of encoding conditions such as varying duration, attention, and word frequency, have not been shown to alter the *Variance-of-targets* parameter as would be expected by this hypothesis (Spanton & Berry, 2020). The *Variance-of-targets* parameter appears to be, in at least some paradigms, highly correlated with recollection (Dede et al., 2013), and thus may have similar relationships to other cognitive variables. Regardless of which model best predicts gF, clarification of the cognitive definition of the variance of distribution parameters (recollection distribution in the CDP and target distribution in the UVSD) is a goal of this dissertation. Due to the lack of cognitive definition of these parameters that has stood up to scrutiny, both were included as predictors, but few hypotheses were made as to what relationships would exist.

In addition to assessing the reliability of the parameter estimates of recognition memory models and examining their ability to predict gF, a third aim of this dissertation was to explore the relationship of each recognition memory parameter to three other cognitive constructs:

processing speed (PS) and verbal fluency., as well as an alternative LTM measure, recall ability A similar approach has been used before in evaluating new cognitive ability constructs. For example, Zerr et al. (2018) used measures of PS, intelligence, and LTM as correlates to help establish the construct validity of a learning efficiency measure. The rationale is that looking at multiple constructs will establish whether a new construct relates to other constructs that, by its nature, it should (convergent validity). For example, learning efficiency should predict school grades, but its relation to certain other constructs (e.g., physical characteristics like height and weight) should be weaker (discriminant validity). Thus, the similarities and differences between the recognition memory models and their relative usefulness in individual differences research can be evaluated based on their parameters' relationships with other select cognitive constructs.

2.3.1 Relation to Recall

Recall memory is a variable that would be expected to be related to at least one memory parameter in each of the recognition memory models because both recall and recognition tests assess episodic LTM. The two tests differ with regard to the demands placed on participants, and according to some models, they involve different types of memory. For example, recall is postulated to be based primarily on recollection (e.g., Mandler, 1980 Quamme et al., 2004). Recollection is defined by retrieval of contextual information and because recall tasks require items must be retrieved in the study context, recall is generally considered a test of recollection. In fact, Mandler argued that the "retrieval processes involved in recognition are essentially the same as those used in recall tasks" (p. 256) . However, few studies have correlated estimates of recollection from a recognition task with recall scores. Theoretically, from the perspective of the DPSD model, recollection should be highly correlated with recall whereas familiarity should be minimally correlated. CDP theorists have made the argument that participants can base recall on familiarity, as participants can recall an item without additional memory of its contextual details (Mickes et al., 2013). This view is not necessarily incompatible with the DPSD model because recall is not a process-pure task (Jacoby, 1991), but a large influence of familiarity on recall is generally more consistent with a model in which the two processes are combined than with a model in which they are independent.

According to the UVSD model, the d' parameter should be related to recall because d' is an index of general memory strength. As long as there exists some general LTM store that links performance between tasks, a measure of an individual's memory strength ought to be related to performance on other LTM tasks. On the other hand, as discussed above, the lack of clear psychological definition of the *Variance-of-targets* parameter leads to a corresponding lack of clear predictions of its relationship with other parameters.

2.3.2 Relation to PS

PS is another variable commonly examined in studies of individual differences in cognitive abilities. PS has been shown repeatedly to be predictive of fluid abilities (for reviews, see Jensen, 2006; Sheppard & Vernon, 2008), and it has been particularly influential in research on children (e.g., Fry & Hale, 1996) and older adults (e.g., Salthouse, 1996) as an ability that can predict performance in a variety of tasks. Particularly relevant to this dissertation, PS has also been shown to be related to general LTM ability (Ghisletta et al., 2012; McCabe et al., 2010). Notably, Salthouse (1996) proposed a PS theory to explain age-related differences in cognitive ability, but the theory also provides an account of mechanisms by which differences in PS can result in differences in memory in individuals regardless of their age. Familiarity has been associated with early processing of study items (Hintzman & Caulton, 1997) and is suggested to be relied on more when participants have to respond quickly (Yonelinas, 2002). Familiarity is partially based on a heuristic whereby participants judge items that are processed easily or faster fluent as more likely to arise from prior study (e.g., Jacoby & Dallas, 1981; Westerman et al., 2003). The role of individual differences is not clearly stated by these descriptions of familiarity, but it is a reasonable idea that fluent processing of old items may be more likely (or the difference in fluency between old and new items larger) for individuals with greater PS, leading to higher familiarity scores for those individuals.

The relation of recollection to PS is less obvious. According to Jacoby (1991), controlled processing such as recollection, unlike automatic processing, is subject to capacity limitations. which are related to PS (Salthouse, 1992). Given that PS is also related to many "higher-order" abilities such as gF (Sheppard & Vernon, 2008), as well as latent LTM variables (McCabe et al., 2010), I hypothesized that faster processing would be related to recollection, and to general memory strength (as characterized by the d' parameter) from the UVSD model. As with the other covariates, the present analysis of the relation of PS to variance parameters (Variance-of-the-remember distribution in the CDP model and Variance-of-lures in the UVSD model) was largely exploratory, as the definition of these variables is unclear.

2.3.3 Relation to Verbal Fluency

Verbal fluency is a construct that represents the ease and speed with which participants can retrieve verbal material. Based on familiarity being described as faster than recollection (Yonelinas, 2002) as well as early descriptions of 'know' responses being related to semantic memory (Tulving, 1985), verbal fluency was hypothesized to be related to familiarity in verbal memory tasks. The relation of verbal fluency to other constructs was hypothesized to be weaker than familiarity. Other researchers have suggested that verbal fluency tasks measure latent controlled search processes and that these processes are related to both recall ability and recognition ability through the shared influence of recollection (Quamme et al., 2004). Therefore, recollection was also hypothesized to be potentially related to verbal fluency. Due to previous associations found between verbal fluency and LTM ability (Hedden et al., 2005), verbal fluency's relationship with other recognition memory parameters, such as d' was also investigated.

Chapter 3: Study 1

This study was designed to establish the reliability of each of the model's parameters. Evaluating the reliability of the parameters is an important prerequisite to considering a model's validity. The reliability of the parameters was estimated by having the same participants perform enough test trials to estimate parameters for each participant twice, and then correlating those estimates. The tests were performed in the same sitting and the parameter estimates were based on different word lists (i.e., alternative-forms reliability).

3.1 Method

3.1.1 Participants

Participants were 68 undergraduate students (48 female) recruited through Washington University in St. Louis' Psychological and Brain Sciences' participant pool who participated online for course credit. Of these participants, 60 identified English as their first language they learned.

3.1.2 Materials

Memory Task. The recognition memory task was based on the remember-know task with confidence ratings used by Rotello et al. (2005) and adopted by Wixted and Mickes (2010) and Ingram et al. (2012). The task was programmed in PsychoPy (Peirce et al., 2019). Words were generated from the English Lexicon Project database (Balota et al., 2007). In all, 480 words were randomly selected from words 4-7 letters in length and were selected to have an above average concreteness rating and frequency; the mean log frequency was 8.85 (Lund & Burgess, 1996). Words were then randomly assigned either to be studied or to serve as lures. There were four study lists each consisting of 60 words presented in either blue or red text, each immediately followed by a test of 60 target words and 60 lures. This is the same list length used by Rotello et al. (2005). In addition to the main study list, two words at the beginning and end of each list served as primacy and recency buffers and were not tested. At test, participants were shown a mix of old and new items and asked for their confidence in their old/new decisions on a 6-point scale using the number keys on their keyboard. Numbers 4-6 were described to participants as indicating increasing confidence that the item was old. If participants made an old (4-6) response, participants were asked whether they "remember" or "know" that they saw the word using the 'R' and 'K' keys on their keyboard.

For the R/K decision, participants were given instructions akin to the "conservative" condition from Rotello et al. (2005), in which they were told that they should only respond "Remember" if they could describe specific contextual or metacognitive details about studying the word. The context (color) was not tested but varying the color of the words during presentation allowed for a context that participants could utilize as a basis for "Remember" judgments in addition to metacognition (e.g., an image that it cued during study). This also kept the task consistent with prior use (Wixted & Mickes, 2010; Ingram et al., 2012).

Detailed instructions for the definition of "Remember" and "Know" were presented on participants' personal computer screen at the beginning of the study, and a brief description was on the top of the screen every time a remember-know judgment was made during the study (i.e., after every old response). Each participant received the same four study-test lists in the same order.

3.1.3 Procedure

Each participant used their own personal computer for this study via an online webpage. After answering a demographics survey, participants read the remember-know instructions. They then took a brief practice test, which consisted of five study items and four test items, in order to familiarize themselves with the interface and the two-step confidence and remember-know judgments. Participants then completed each of the four study-test lists. Each target word was presented on the screen for 1.75s with a .25s inter-stimulus-interval and participants were instructed to try to remember each word for a later test. The participants then took the test, which was self-paced. After each test, participants were informed how many more memory tests were remaining and that they could take a break if they desired. The study was self-paced and there was no time limit for completion.

3.1.4 Analysis

Three different sets of parameters (one set for each model) were estimated for each participant two times. Due to the number of responses estimated to be necessary for reliable parameters estimates, two of the 120 item test scores were combined for each parameter estimate. Responses to the first two recognition tests and responses from the second two recognition tests were combined and these data were used to fit each model and derive parameter estimates. This meant that each estimate was based on 240 total responses to 120 targets and 120 lures. Performance from tests 1 and 2 is referred to as Part A performance and performance from tests 3 and 4 as Part B performance. Of interest was simply the correlation between the same parameter's estimates based on Part A and Part B performance for each individual. The Intraclass Correlation Coefficient (ICC) is reported as the main test of each parameter's reliability. The ICC used was the two-way random effects model for absolute agreement, which corresponds to ICC (2,1) in Shrout and Fleiss' (1979) nomenclature, and was calculated using the 'psych' package in *R* (Revelle, 2020).

Unequal Variance Model. The UVSD model was fit based on the ROC from the confidence ratings, ignoring the R/K decision. Data were fit using the MATLAB tool designed by Koen et al. (2016). Maximum likelihood estimation was used to estimate the best-fitting estimates of the data based on the model's free parameters. For the UVSD, the free parameters of interest are discriminability (d') and variance (V_{old}) of the target distribution. Five confidence criteria are also estimated parameters in the model (one less than the number of confidence levels in the paradigm) for a total of seven free parameters.

Dual Process Signal Detection Model. The DPSD could theoretically be fit by the remember/know decisions or the ROC, but not both simultaneously. One test of this model came from estimates based on the R/K decisions. Recollection was equal to the rate of remember responses to target items (R_{old}) minus the rate of remember responses to lures (R_{new}). Familiarity was the rate of know response to old items that were not remember ($F_{old} = K_{old} / 1 - R_{old}$) minus the rate of know responses to new items corrected ($F_{new} = F_{new} / 1 - R_{new}$).

As an additional test of the model, the ROC method was used to calculate separate DPSD parameters using the Koen et al. (2016) MATLAB tool. The parameters for this version of the model are estimates of recollection and familiarity, plus five confidence criteria, for a total of seven free parameters.

Continuous Dual Process Model. The CDP was fit based on combination of confidence and remember/know judgments using the same MATLAB code used by Wixted and Mickes (2012). As in the ROC Toolbox, the model was fit using maximum-likelihood estimation. The main parameters of interest were the mean of the recollection (*mean-of-R*) and familiarity (*mean-of-F*) distributions, as well the standard deviation of the recollection distribution (*SD-of-R*). In addition, there were five confidence criteria plus a remember criterion for a total of nine free parameters.

3.2 Results

All participants performed above chance, as indicated by a Corrected HR (Hit Rate minus False Alarm Rate) greater than 0, on both Part A and B. In addition, an *a priori* inclusion criterion was that all participants must have at least 6 hits on each part, as it gave the opportunity to record at least one 'Remember' and 'Know' response at each confidence level of old (4-6). No participants were removed based on this criterion. Welch's unequal variances t-test indicated no differences between native English speakers and non-native speakers in Corrected Hit Rate (HR) for either memory list A, $t(10.62) = 0.33$, $p = .750$ or memory list B, $t(14.48) = 1.15$, $p = .271$. Given that all participants were enrolled full time in courses taught in English at a competitive university, all reported analyses are for the entire sample. When the analyses were repeated with only native English speakers, the results did not differ (see Appendix A).

Table 3.1 Descriptive statistics and reliability for recognition task

Measure	Mean	SD	Skew	Kurtosis	Pearson *r* (95% CI)	ICC (95% CI)
HR Part A	.60	.20	-0.39	-0.23	.83	.82
HR Part B	.56	.24	-0.05	-0.70	(.79, .91)	(.74, .88)
FAR Part A	.19	.14	0.48	-1.12	.84	.84
FAR Part B	.18	.16	1.01	0.27	(.73, .90)	(.77, .89)

Note. HR = Hit Rate, FAR = False Alarm rate. 95% confidence intervals in parentheses. Commonly used interpretations of ICC values are: excellent (.8), good/substantial (.6), and moderate (.4) levels of reliability (Cicchetti & Sparrow, 1981).

Descriptive statistics for the overall recognition memory performance are provided in Table 3.1. As can be seen, the mean performance, characterized by hit and false alarm rates, was relatively similar across the two parts and performance was highly correlated across the two parts. In addition, the difference between the two, the Corrected HR, was highly reliable across the two parts (*ICC* = .82, 95% CI [.74, .88]).

3.2.1 Reliability of Models

Inspection of the distributions for the CDP and UVSD parameters suggested that several participants' parameter estimates were outliers (e.g., a d' estimate greater than 100). Extreme outliers were removed from the data. Extreme outliers were defined as any value that was more than 3 times the Interquartile range (IQR) above the third quartile (i.e., Q3 + 3 IQR) or below the first quartile (i.e., Q1 − 3 IQR). In an effort to keep the maximum possible sample size, participant's data were only removed for a given model if one of the estimated parameter values for that particular model was an extreme outlier (i.e., pairwise deletion). This led to the possibility of the sample size being greater for some analyses than others. In practice, there were no extreme outliers for the DPSD model, whereas there were six each for the CDP model and the UVSD model. The descriptive statistics and reliability estimates for the memory parameters from the DPSD, the CDP, and the UVSD are presented in Table 3.2. The reliabilities for the non-memory parameters (i.e., criteria) are presented in Appendix A. Both the DPSD parameters, recollection (ICC = .84) and familiarity (ICC = .78), exhibited reliability around .80, defined as strong reliability by commonly used criteria (Cicchetti & Sparrow, 1981). The d' parameter from the UVSD model also exhibited strong reliability (ICC =.80). The V_{old} parameter, on the other hand, was below the standard of good reliability (.60) but above moderate reliability (.40). None of the parameters from the CDP model exhibited strong reliability. Both the *mean-of-R* (ICC = .54) and the *SD-of-R* (ICC = .58) fell just below the threshold for good reliability. Of the most concern for the CDP model, the reliability of *mean-of-F* was weak (.23), and the lower bound of the 95% Confidence Interval was only slightly above zero.

Table 3.2 Descriptive statistics and reliability for model parameters with extreme outliers removed

Parameter	*n*	mean	SD	skew	kurtosis	Pearson *r*	ICC
DPSD model	68						
Recollection A		0.21	0.19	0.92	0.23	.85	.84
Recollection B		0.21	0.23	1.15	0.39	(.76, .92)	(.77, .89)
Familiarity A		0.30	0.21	0.60	0.14	.80	.78
Familiarity B		0.27	0.25	0.79	-0.15	(.68, .88)	(.69, .85)
CDP model	62						
Mean-of-R A		1.15	1.84	0.41	0.55	.54	.54
Mean-of-R A		1.36	1.73	0.56	0.18	(.30, .71)	(.37, .67)
Mean-of-F A		0.87	1.17	-0.01	1.37	.23	.23
Mean-of-F B		0.52	1.30	0.32	-0.28	(-0.05, .50)	(.02, .42)
SD-of-R A		1.74	0.74	0.38	1.61	.59	.58
SD-of-R B		1.73	0.93	0.54	0.79	(.39, .74)	(.43, .70)
UVSD model	62						
d′ A		1.44	1.05	1.31	1.79	.83	.80
d′ B		1.42	1.36	1.60	2.83	(.66, .92)	(.71, .97)
V_{old} A		1.41	0.38	1.48	2.81	.55	.54
V_{old} B		1.43	0.48	0.56	0.57	(.32, .72)	(.37, .67)

Note. 95% confidence intervals in parentheses. DPSD recollection and familiarity scores are based on estimates directly from proportion of ‘remember’ and ‘know” responses.

A potential concern was that the lack of reliability of the parameters from the CDP might be due in part to some of the participants not having enough data for both remember and know at each level of confidence to allow accurate parameter estimates. In order to assess this possibility, both the data for any participant who did not have at least 3 remember and 3 know responses to old items on each list and the data for any participant that did not use the full confidence scale (at least one 1-6 ratings) were deleted listwise. Applying these criteria caused the removal of 17 participants’ data, reducing the sample size to 51 participants. The same correlations were calculated and the results are reported in Table 3.3.

As can be seen in Table 3.3, the reliabilities for the DPSD and UVSD parameters were similar to those previously found after removing outliers directly. However, the CDP's *SD-of-R* exhibited only an ICC of only .10. Visual inspection of the distributions of the *SD-of-R* parameter indicated that this was influenced by outliers. As with the analysis of the full data and the data with outliers removed, the mean of F was not significantly correlated across the two parts. Therefore, regardless of whether outlying scores are removed or not, the two parameter estimates show little to no relationship**.**

Table 3.3 Descriptive statistics and reliability for model parameters with 51 participants

Parameter	mean	SD	skew	kurtosis	Pearson *r*	ICC
DPSD model						
Recollection A	0.26	0.17	0.92	0.36	0.81	0.78
Recollection B	0.27	0.23	0.92	-0.13	(.66-.90)	(.67, .86)
Familiarity A	0.34	0.21	0.61	0.04	.81	0.81
Familiarity B	0.31	0.24	0.86	-0.1	(.66-.90)	(.71, .87)
CDP model						
Mean-of-R A	1.66	1.7	0.49	0.19	.53	.48
Mean-of-R B	1.66	2.71	-0.65	4.93	(.30-.74)	(.28, .64)
Mean-of-F A	0.91	1.09	0.91	2.85	.14	.13
Mean-of-F B	0.7	1.68	1.63	3.67	(-.17-.41)	(-.10, .35)
SD-of-R A	2.09	1.41	2.6	8.77	.14	.10
SD-of-R B	2.58	3.02	4.84	26.34	(.02-.56)	(-.13, .33)
UVSD model						
d' A	1.79	1.38	1.65	2.49	.78	.72
d' B	1.82	2.04	2.31	8.3	(.64-.90)	(.59, .82)
V_{old} A	1.55	0.68	3.49	15.36	.43	.43
V_{old} B	1.62	0.78	1.94	5.61	(.26-.73)	(.22, .60)

Note. 95% confidence intervals in parentheses. DPSD recollection and familiarity scores are based on estimates directly from proportion of 'remember' and 'know" responses.

3.2.2 DPSD from Confidence

As mentioned, the DPSD model parameters can be calculated both from the remember and know responses and from the ROC data. The DPSD model from ROC data are reported in Table 3.4. The same procedures were used with regard to outliers and only one extreme outlier was detected, and the reliability of the parameters was relatively unchanged by the exclusion of this outlier. Overall, the reliability of both the recollection parameter (ICC = .75) and the reliability of the familiarity parameter (ICC = .67) were relatively high although they were both slightly below the proposed cutoffs for strong reliability.

Table 3.4 Descriptive statistics for DPSD model based on confidence judgements.

Parameter	*n*	mean	SD	skew	kurtosis	Pearson *r*	ICC
Full Sample	67						
Recollection A		0.25	0.20	1.20	1.43	.75	.75
Recollection B		0.25	0.23	1.05	0.27	(.57, .86)	(.64,.82)
Familiarity A		0.69	0.53	0.64	0.3	.70	.67
Familiarity B		1.15	1.84	0.41	0.55	(.56, .85)	(.54, .77)
Partial Sample	51						
Recollection A		0.30	0.22	0.95	0.50	.74	.74
Recollection B		0.29	0.24	0.87	-.30	(.57, .86)	(.61,.83)
Familiarity A		0.82	0.56	0.63	0.35	.72	.67
Familiarity B		0.87	0.86	1.78	3.57	(.54, .87)	(.54,.77)

Note. 95% confidence intervals in parentheses.

3.2.3 Prediction errors of models

An additional analysis was conducted to determine if the reliability differences between the parameters of the different models had any cost in terms of the models' cross-validation across two datasets (the two parts of the study). If a model is overfitting to specific data, this may have costs in terms of its ability to predict a new data, even if the correlations in performance were high across she two parts in this study. Traditionally, the closeness of fit of each recognition memory model is tested using the ROC data containing old and new responses (and for the CDP the ‘R’ and ‘K’ responses) at each confidence level. In order to better understand the overall reliability of the models, this analysis tested the ability of the model parameters to predict performance in the other half of the study (i.e., parameters from part A predicted performance on part B and vice-versa). Rather than assessing the old and new responses at each level of confidence, the data were the participants’ HR and FAR, chosen as indicators of overall performance.

Each model’s best-fit parameter estimates for a participant were used to calculate model estimates of HR and FAR for each participant. The absolute values of the difference between these estimates and the observed HR and FAR for the participant on the opposite half of the study are reported in Table 3.5. The DPSD estimates based on the R/K data were not used for this analysis because the estimates correspond exactly to the observed HR and FAR.

Table 3.5 Absolute Vales of difference of observed versus predicted Hit and False Alarm Rates

model	Hit Rate difference			False Alarm Rate difference		
	mean	SD	max	mean	SD	max
Part A predicted minus observed Part B						
UVSD	0.10	0.08	0.3	0.08	0.09	0.46
DPSD	0.10	0.07	0.28	0.08	0.09	0.48
CDP	0.11	0.08	0.29	0.07	0.07	0.29
Part B predicted minus observed Part A						
UVSD	0.10	0.08	0.32	0.08	0.08	0.43
DPSD	0.06	0.05	0.19	0.07	0.08	0.43
CDP	0.12	0.12	0.68	0.04	0.07	0.39

Note. Each model’s predicted HR and FAR were calculated for Part A and Part B separately, then subtracted from the observed HR and FAR for the other half. DPSD is based on ROC version of that model.

Repeated measures analysis of variance (ANOVA) was conducted on the difference scores presented in Table 3.5, separately for the HRs and the FARs. Data were collapsed across the two predictions (estimates from A predicting performance on B and vice-versa). For the Hit Rates, Mauchly’s test of sphericity indicated the sphericity assumption was violated, $W = .65$, p

< .001, and therefore, a Greenhouse-Geisser correction was used. There was a significant main effect of model, $F(1.44, 87.55) = 14.00$, $MSE = 0.002$, $p < .001$, $\eta_p^2 = .04$. Post-hoc t-tests (corrected for multiple comparisons using the Holm-Bonferroni correction) indicated that the absolute vale of the differences scores was significantly smaller for the DPSD ($M = .08$) compared to the CDP ($M = .11$), $t(118) = 5.14$, $p < .001$, and the UVSD model ($M = .10$), $t(118) = 1.49$, $p = .001$. The CDP and the UVSD did not differ significantly, $t(118) = 1.44$, $p = .138$.

For the FARs, Mauchly's test of sphericity indicated the sphericity assumption was violated, W = .10, p < .001, and therefore, a Greenhouse-Geisser correction was used. A one-way repeated measures ANOVA found a significant main effect of model, $F(1.05, 62.36) = 17.58$, ,$MSE = 0.001$, $p < .001$, $\eta_p^2 = .02$. Post-hoc t-tests (corrected for multiple comparisons using the Holm-Bonferroni correction) indicated that the prediction errors for the CDP ($M = .06$) were significantly smaller compared to the DPSD ($M = .08$), $t(118) = 5.21$, $p < .001$, and the UVSD ($M = .08$), $t(118) = 5.06$, $p < .001$. The DPSD and the UVSD models' prediction errors did not differ significantly, $t(118) = 0.14$, $p = .877$.

3.3 Discussion

The analysis of the reliability of the parameters from the recognition memory models led to a few clear conclusions. The recollection and familiarity parameters of the DPSD model based on remember and know judgments were highly reliable. The DPSD model's parameters can also be calculated based on the ROC data, and estimates from this model also exhibited substantial reliability. For the UVSD model, the *d′* parameter was highly reliable, and the variance-of-targets parameter was moderately reliable. In contrast, the familiarity parameter from the CDP, *mean-of-F*, was not reliable, regardless of whether outliers were included in the analysis. The other memory parameters for the CDP (i.e., the *mean-of-R* and the *SD-of-R*) were shown to be moderately reliable.

Comparing the models, parameters of the DPSD model are the most reliable. The R/K estimates were clearly in the “good” to borderline “excellent” range for estimates of reliability (Cicchetti & Sparrow, 1981; Landis & Koch, 1977) and it was the only model that had at least good reliability for all of its parameters. Furthermore, the DPSD model estimated from the ROC data also produced estimates of familiarity and recollection to old parameters with good reliability. This suggests that there are two ways that recollection and familiarity can be calculated and still be reliable.

An additional factor in assessing the DPSD model is that the task used was somewhat inconsistent with the model itself. The DPSD model suggests that confidence ratings and remember and know judgments are independently sufficient to estimate recollection and familiarity. According to the DPSD, recollection should be associated with high confidence, asking for confidence in addition to remember judgments should not be necessary. Theoretically, from the perspective of the DPSD, participants would be inclined to rate all above-threshold recollections with remember judgments and the highest confidence rating (6). However, due to task demands, participants may not respond with the highest confidence for all remember judgments when those judgments are asked to be made separately, since they may assume that the researcher wants those judgments to differ. Therefore, it is possible that the reliability estimates found in the study, although strong, may be lower estimates compared to what would be found if participants were asked to only make either a remember- know judgment or confidence ratings and not both.

As in previous studies (e.g., Rotello et al., 2005), when asked to make both remember/know judgments and confidence ratings, participants did not exclusively use the highest confidence rating for remember judgments, as would be expected from a strict DPSD perspective. The correlation between the two DPSD estimates was $r = .67$ for recollection and $r = .85$ for familiarity for Part A and $r = .65$ and $r = .87$ for recollection and familiarity respectively for Part B.

The UVSD model produced parameter estimates that were relatively reliable across Part A and B for participants. The d' parameter exhibited excellent reliability, but the Variance of targets (V_{old}) parameter was only moderately reliable. Therefore, although the d' value is a strongly reliable metric of performance, the V_{old} parameter may be picking up on noise as much as from a meaningful behavioral signal. If this is the case, the UVSD model may fit an individual participants' data better than the DPSD model on a single recognition test (Heathcote, 2003; Healy et al., 2005), in part because it is overfitting their data. This prediction was supported by the analysis of prediction errors, which showed that the UVSD had significantly more errors when predicting HR on the alternative part than the DPSD. It also had greater error than the CDP at predicting FAR. The following study (Study 2) was designed to inform whether the V_{old} parameter is providing meaningful insight into cognition by assessing this parameter's relationship with other cognitive variables. In addition, comparing d' to recollection and familiarity, will help in understanding whether a single memory parameter captures all the meaningful variance between individuals' performance on recognition tests.

With respect to the CDP model, the lack of reliability of the familiarity (*mean-of-F*) parameter in the CDP model is highly problematic for the model. Even the most generous interpretation would put the upper confidence interval at a barely acceptable level (upper confidence limit = .42). Furthermore, both the *mean-of-R* and *SD-of-R* parameters exhibited significantly worse reliability than the recollection parameter from the DPSD model (based on remember-know responses), as evidenced by a lack of overlap of the confidence intervals. These results are particularly a problem because the recognition memory task used in this study was specifically chosen because it is the only task that researchers have used to estimate CDP parameters. Therefore, if estimates from this task do not produce reliable estimates of familiarity for an individual, it is not possible to get reliable estimates of familiarity based on the CDP model based on extant recognition memory paradigms. The only recourse is to either design yet another new recognition memory task or to try to increase the number of trials to see if that impacts the reliability. Considering that this study's design already had 240 memory judgments per individual for each parameter estimate, it is unclear that more trials would improve reliability. Furthermore, not only was participants' overall performance reliable, as characterized by their hit and false alarm rates (Table 3.1), so were the parameter estimates for the other models. The results of this study suggest that the CDP is overparameterized, given that two of its parameters showed some evidence of reliability but the third memory parameter did not. This overfitting was supported by the fact that the CDP's predicted HRs were further from the observed HR in the other half of the study than the DPSD despite having more parameters. However, the model's additional parameters may benefit its ability to predict FAR, as it outperformed the other two models. In contrast, the DPSD model and UVSD model, which each have two fewer free parameters than the CDP, exhibited more reliable parameter estimates for an individual.

In conclusion, the results of this study indicate that the DPSD model parameters are more reliable than those of the UVSD and that the CDP model's parameter are the least reliable. On the one hand, the UVSD model's parameters have moderate to borderline "excellent" reliability, and the following study will determine if those parameters are more useful or less useful than the DPSD parameters for predicting other cognitive abilities. On the other hand, the CDP model's measure of familiarity, the *mean-of-F* parameter, did not demonstrate even moderate reliability. Thus, this measure of familiarity is not useful for understanding individual differences. The following study will determine whether the other CDP parameters, the estimates of the mean and the *SD-of-R*, which were moderately reliable, are useful for describing individual differences when compared with the estimates from the DPSD. Unless the *mean-of-R* and the *SD*-of-R parameters are capturing some aspect of recollection that the DPSD model does not, the CDP model does not appear to be a more useful dual-process model than the DPSD. It would therefore be reasonable to conclude that the CDP is needlessly overparameterized, as demonstrated by the lack of reliability of the *mean-of-F* parameter.

Chapter 4: Study 2

The previous study evaluated the reliability of the model parameters for the DPSD, CDP, and UVSD models of recognition memory. Each of the models was determined to have at least one reliable parameter and a second parameter with at least some degree of reliability. Reliability is an important and necessary attribute of any individual difference variable, but a reliable construct is not necessarily valid. One way in which the recognition memory models' validity can be assessed is based on whether the parameters ought to relate to established cognitive variables that are theoretically related, termed convergent validity. Fluid reasoning (gF) should rely in part on memory abilities and has been shown to be related to LTM ability (e.g., Unsworth, 2019). As discussed in Chapter 2, recognition memory models should predict an individual's fluid reasoning. In addition, at least one of the model parameters should be related to an individual's recall ability, because although recognition and recall are different tasks, they should rely on at least some of the same general memory abilities. In the following study these constructs were assessed with multiple measures and then compared to the recognition memory model parameters with Structural Equation Modeling (SEM).

In addition to assessing each model's overall convergent validity, the study was designed to provide a better description of each of the model's parameters. For example, processing speed (PS) and being able to come up with words easily (verbal fluency) may be related to a person's ability to make recognition memory judgments using familiarity. Evaluating each of the models' parameters' relationship with recall and gF also provided a better understanding of model parameters that have not always been well-defined cognitively. By directly estimating the parameters this offers a novel and perhaps more accurate measure of the processes compared to previous studies of individual differences in recognition memory. The specific hypotheses and rationale for the relationships between the recognition parameters and the other cognitive variables were outlined in Chapter 2.

Finally, the analyses were designed to provide a comparison of the models. If a model does not exhibit convergent validity, this would be evidence against the model. In the case that each model exhibited at least moderate convergence with these constructs, the models were compared for their ability to predict the constructs, to help determine which is the most useful model of individual differences in recognition memory.

4.1 Method

4.1.1 Participants

Participants were 220 Washington University undergraduate students recruited through the university's Psychology department subject pool, with the aim of having 200 participants with complete data. Ten participants either did not take the second part of the study or did not follow instructions on one part of the study. An additional three participants' data were lost for the second part of the study due to malfunctioning of the web server. Four participants' recognition data were unusable because they only used 1 or 2 ratings on the 6-point confidence scale. This left a total of 201 participants (M_{age} = 20.1 years, age range: 18-22 years) with complete data. Of the 201 participants in the final sample, 21 participants reported that English was not one of the first languages they learned. In addition, 163 participants (80%) were female and the sample was majority non-Hispanic White (57%).

4.1.2 Measures

Long Term Memory Tasks

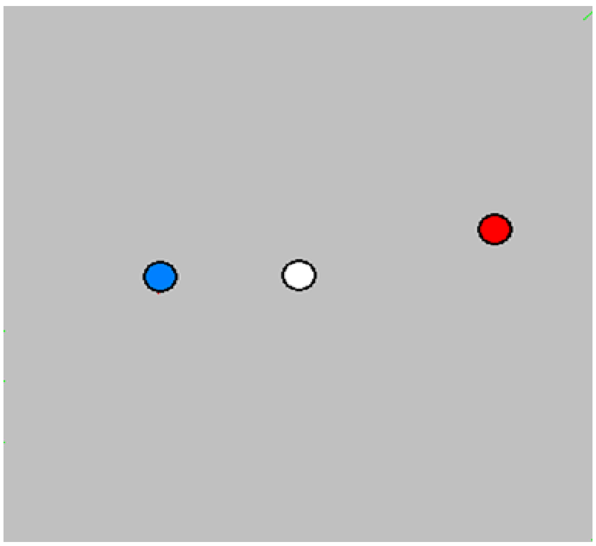
The recognition memory task was the same R/K with confidence task described for the previous study. Only the first two words lists from the previous study were used. As in the previous study, performance from the two lists were combined and the 240 judgments were used to attain parameter estimates for each recognition memory model. The only new element for this study was that a recall task was added to the end of each list. Immediately after completion of the entire R/K recognition test, participants were asked to recall any items that they could from the study list.

An additional recall task of a new list of 25 words occurred. For the recall list, the 25 words were randomly chosen from the list of words with the same characteristics described for the recognition memory tasks. The words were presented for 2 s in the middle of the screen, and after a 30 s distractor task in which participants verified true/false arithmetic equations, participants were asked to recall as many words as possible. Scores for the recall test were proportion of words correctly recalled. The scores for the R/K list were divided by 25 instead of the maximum possible score of 60 studied words, to keep the scale consistent between tests (no participant correctly recalled more than 25 words).

PS Tasks

Distance Judgment. For each trial, a white dot appeared in the center of the screen with a red dot to the right and a blue dot to the left. Participants were asked to decide whether the left or the right dot was closest to the central dot by pressing the 'z' key for left and the '/' key for right. A sample trial is shown in Figure 4.1. Participants performed 10 practice trials with feedback on whether they answered correctly (i.e., chose the closer dot), followed by 22 test trials. The first 2 test trials served as buffers and were not included in the analysis. For all processing speed tasks, response time (RT) and accuracy were recorded, with a participants score on the task calculated as the mean RT on correct trials.

A



B

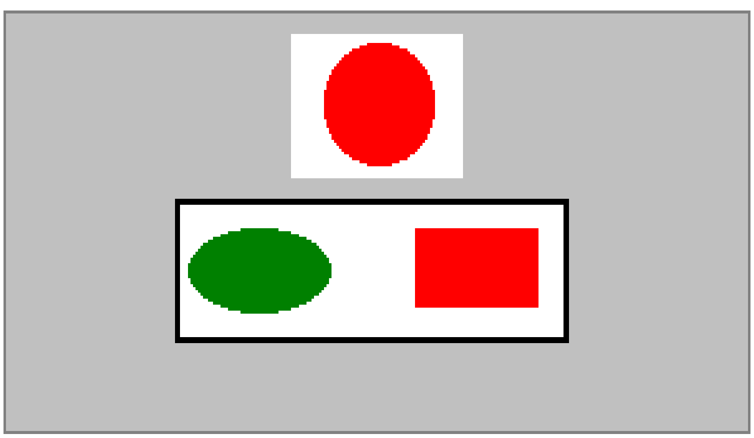


Figure 4.1. Sample Trials for Processing Speed Tasks A: Sample dot task trial. The correct response would be the left key as it the left is closest to the center dot. B: Sample Shape task trial. The correct response would be the left key as the shape matches the sample shape (top center).

Shape Judgment. Participants viewed a sample shape (e.g., a circle) and two choice shapes (e.g., an oval and a rectangle) that appeared side-by-side below the sample shape. Participants decide which of the two choice shapes is more similar to the sample shape by pressing the ‘z’ key for the shape on the left and the ‘/’ key for the shape on the right. A sample trial is shown in Figure 4.1. Participants performed 10 practice trials with feedback followed by 22 test trials, with the first 2 test trials serving as buffers.

Animal Categorization. Participants were presented with words one at a time on the screen. Each word was the name of either an animal or a fruit/vegetable. Participants were asked to respond whether each word is the name of an animal by pressing the ‘z’ key for animal and the ‘/’ key for a non-animal. Participants performed 10 practice trials with feedback followed by 42 test trials, with the first 2 test trials serving as buffers.

Fluid Reasoning Tasks

Raven’s Advanced Progressive Matrices (RAPM*)*. In the RAPM, each problem contained a 3x3 matrix with one element missing and eight elements displayed (Raven et al., 1998). Participants were asked to select, from six options, the element that completed the matrix along both the rows and columns. The half-set was given, which contains 18 test trials. After being shown instructions and a sample problem, participants were given a maximum of10 minutes to complete as many of the test trials as possible. Scores were proportion of test trials solved correctly.

Number Series. The Number Series task consists of 15 items (Thurstone, 1938). In this task, subjects see a sequence of numbers that follow a logical pattern. The participant’s task was to choose, from five available options, the next number in the sequence. Participants performed four practice problems and were then allowed a maximum of 5 minutes to complete as many of the 15 test items as possible. Scores were proportion of test trials solved correctly.

Cattell Culture Fair Matrix Reasoning. Like the RAPM, the Cattell Culture Fair Matrix Reasoning (Cattell, 1973) is a visual reasoning task in which a matrix of objects is displayed with the last item missing. The participant's task was to select, from five available options, the image that completed the sequence. Participants were allowed a maximum of 10 minutes to complete all 20 problems. Scores were proportion of test trials solved correctly.

Fluency Tasks

Category. Participants were asked to type every word they could think of in a single category in 60 s. Two category lists (fruits and furniture) were used. Scores for all fluency tasks were number of words produced per second (total words divided by 60). Words were determined to be acceptable part of the fruit category if they were non-repeated names of a fruit either botanically and/or culinarily. Words were judged to be valid furniture if they could reasonably be considered a piece of furniture; due to the lack of specificity in instructions, common appliances or similar home goods were accepted as judged by a single rater. Any misspellings that were within one letter of a correct word were accepted.

Letter. Participants were asked to type every word starting with a particular letter in 60 seconds. Two letter lists (M and S) were used. All words starting with the correct letter that appeared in an English dictionary were accepted.

4.1.3 Procedure

The study occurred online across two sessions lasting approximately 25-30 minutes each. The participants competed all the tasks in the same order, as is recommended in studies of individual differences (Goodhew & Edwards, 2019). During Session 1, participants completed both lists of the R/K with confidence recognition task. The standalone 25-word recall task occurred after both study-test lists were completed. The PS tasks occurred between memory tasks. The complete order of tasks was as follows: R/K List 1, recall of List 1, dot task, R/K List 2, recall of List 2, animal categorization, 25-word recall, shape judgment. After completion of

Session 1, participants were sent a link to Session 2. They could complete Session 2 any time during the week that followed. During Session 2, the fluid reasoning tasks were completed, with the fluency tasks occurring between. The session began with the fluency task asking to list words starting with the letter “S”, followed by Raven’s, fruit fluency, Number Series, “M” fluency, Cattell, and furniture fluency.

4.1.4 Analysis

The primary analysis for this study used SEM techniques. The general SEM approach used a two-step procedure for each model (Anderson & Gerbing, 1988). The measurement (CFA) models were first fit to the data, and any adjustments were made if the fit was poor. Then, the structural regressions models were fit and any alternative models (e.g., setting non-significant paths to zero) were compared. All analyses were conducted in R statistical software (R Core Team, 2018) and the R package *lavaan* (Rosseel, 2012) was used to fit all SEM models to the variance-covariance matrices. The model fits reported were the model Chi-squared, the Root Mean Square Error of Approximation (RMSEA; Steiger, 1990), Comparative Fit Index (CFI; Bentler, 1990), and Standardized Root Mean Square Residual (SRMR), based on the recommendations of Kline (2016), as well as the AIC and BIC. A non-significant Chi-square test indicates a good fitting model, as does an RMSEA < .05, a CFI >.95, and an SRMR < .08. Confidence intervals around the RMSEA were also reported and in an ideal fitting model would not exceed .08. There were two main analyses, concerning the relationship of the recognition memory decision models’ parameters with gF and recall. Relationships with PS and fluency were also assessed.

gF

The first analysis assessed the ability of the parameters from each recognition memory model to predict gF. To assess this relationship, a latent factor gF was modeled by performance on the three gF tasks. Each model allowed for two recognition memory parameters to predict gF. Because only a single estimate of each parameter was calculated per individual, each was treated as a single factor latent variable. Using the reliability estimates from Study 2, error variances were estimated by subtracting the reliability from 1.0 and multiplying that by the variance: $(1-r_{aa})* s^2$ (Kline, 2016). PS was also allowed to predict gF as PS has been shown to be related to gF (see Sheppard & Vernon, 2008, for a review). Recognition memory was hypothesized to mediate the relationship between PS and gF. A sample mediation path model is shown in Figure 4.2. Of interest is the path from each recognition parameter to gF. The secondary outcome of interest was the relationship with the parameters to PS, and how much if at all, they mediated the relationship of PS with gF.

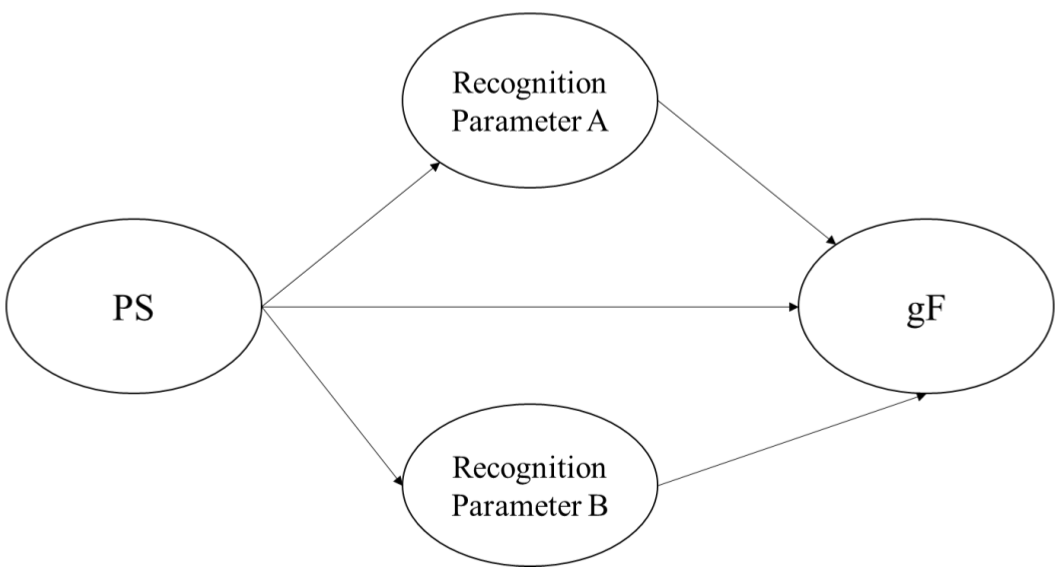


Figure 4.2. Path model with all hypothesized potential paths. Recognition parameter A and Recognition Parameter B are single-factor latent variables based on parameters from each model (e.g., recollection and familiarity). PS and gF are latent variables based on the three measures of each ability.

For both estimates of the DPSD model, the two parameters included in the model were recollection and familiarity. For the UVSD model, discriminability (d') and variance of targets (V_{old}) were allowed to predict gF. For the CDP model, mean of the recollection distribution (*mean-of-R)* and *SD* of recollection distribution (*SD-of-R)* were allowed to predict gF. Based on the results of the previous study, mean of familiarity (*mean-of-F)* was not included in the CDP model because it was not reliable. In addition to establishing convergent validity through assessing relationship between recognition parameters and gF, the models were compared. Given an adequately close fit of each model, the primary way the models were compared was based on the proportion of variance of gF explained by each model's parameters. Bias-corrected and accelerated (BCa) bootstrapped confidence intervals were calculated based on 5,000 bootstrap samples to determine upper and lower bounds of both the overall model and each parameter's influence on gF, to determine if there was any evidence of one model predicting more variance in gF than the others.

Recall

The analysis of recall evaluated the fit of single-factor CFA models. For each of the CFA models, the scores from the three recall tasks loaded on the single factor plus a fourth indicator, which was a parameter from one of the 3 recognition memory models. As long as the model was a good fit to the data, the strength of the recognition memory parameter's relationship with the factor, characterized by the standardized loading (structural coefficient), was assessed. In addition to the standardized loading, evaluation of the models was assessed through composite reliability of the factor based on the ratio of the explained variance versus the total variance (Bollen, 1980) and the average variance explained (AVE) in the indicator variables. These measures helped to form an overall idea of how well the factor was explaining variance in recall ability and the recognition parameter of interest.

For both versions of the DPSD, recollection was allowed to load on the recall factor. For the CDP, the *mean-of-R* was the primary indicator of recall. A separate model also tested the relation of *SD-of-R* to recall because this parameter was potentially related as well. For the UVSD, *d′* was the main indicator assessed. However, variance of targets (V_{old}) has been suggested to be similar to recollection (Dede et al., 2013), so both parameters could be potentially related to recall. Independent models were fit to assess the loading of each parameter on the recall factor. CDP theorists have argued that participants can base recall on familiarity (Mickes et al., 2013) and this has been suggested to be consistent with the DPSD as well (Yonelinas & Parks, 2007). Although the CDP model did not have a reliable familiarity parameter, the familiarity parameter from the DPSD's relationship with recall was tested by forming a model that assessed the latent correlation between a single-indicator familiarity latent factor and the recall factor.

In addition to fitting each recognition memory model independently, an additional analysis compared the fit of the best indicator of recall from each recognition memory model by including them all in the same CFA model. The recognition memory parameters were standardized so that their scales were equivalent. The overall fit of a single-factor model with each of the recognition memory parameters and the three recall tests loading on the single-factor was first assessed. Then, the constraint that each of the loadings of the recognition memory parameters was equal was added. Additional models were then tested, so that one of the recognition memory parameters was allowed to be freely estimated, while the other two parameters were constrained to be equal. Whether any of the models proved to be a better fit than the model that constrained the parameters to be equal would determine whether the association with the recall factor was different for any of the parameters.

Fluency

A final analysis involved the relationship of fluency with specific model parameters. In particular, familiarity's relationship to this construct was of interest. A verbal fluency latent factor was formed based on the four fluency tasks. Latent correlations with the verbal fluency factor were tested for several variables. As in the analysis of gF, a single-indicator familiarity latent factor was formed and its correlation with the verbal fluency factor was assessed. Due to the similarity of UVSD d' and familiarity, the correlation of d' with verbal fluency was also assessed.

4.2 Results

Table 4.1 Totals and proportions of memory responses by confidence

Items	Confidence	Remember	Know	New
Old	6	5062 (.19)	2668 (.10)	-
	5	650 (.02)	3107 (.12)	-
	4	324 (.01)	3540 (.14)	-
	3	-	-	4929 (.19)
	2	-	-	3686 (.14)
	1	-	-	2194 (.08)
New	6	440 (.02)	453 (.02)	-
	5	297 (.01)	1395 (.05)	-
	4	303 (.01)	2849 (.11)	-
	3	-	-	7494 (.29)
	2	-	-	7203 (.28)
	1	-	-	5726 (.22)

Note. Numbers in parentheses are proportions of responses for each item (e.g., the proportion of Old items receiving a 6 confidence and Remember response was .19)

The criteria that were established for the previous study in identifying extreme outliers for the recognition memory parameters were applied. Based on these criteria 9 participants' data were determined to contain extreme outliers in their recognition memory data and were removed listwise. This left a total sample of 192 participants. The overall pattern of results was unchanged when these participants' data were included. The proportions of responses to targets and lures separated by level of confidence and R/K decision during the recognition memory task are presented in Table 4.1 Across participants and confidence levels, the false alarm rate was .22, and the hit rate was relatively low, .58. In addition to the low overall hit rate, the proportion of remember responses was relatively low, with only 22% of old items receiving "R" responses, most of which were accompanied with a confidence rating of 6.

Table 4.2 Descriptive statistics for all tasks and recognition memory parameters

Variable	Mean	SD	Median	Skew	Kurtosis
Recollect R/K	0.19	0.16	0.17	0.79	0.03
Famil R/K	0.29	0.14	0.29	0.03	-0.64
Mean-of-R CDP	1.32	1.35	1.10	0.29	-0.52
Mean-of-F CDP	0.64	1.07	0.67	-0.02	0.17
SD-of-R CDP	1.88	0.75	1.82	-0.07	0.85
d' UVSD	1.32	0.73	1.23	0.70	0.61
V_{old} UVSD	1.46	0.34	1.39	1.21	1.72
RT dot task	0.64	0.12	0.62	0.77	0.58
RT animal task	0.64	0.09	0.62	0.72	0.12
RT shape task	0.59	0.10	0.57	0.83	0.51
RAPM	0.44	0.19	0.50	-0.31	-0.43
Number Series	0.66	0.18	0.67	-0.46	-0.49
Cattell	0.39	0.13	0.40	-0.06	0.13
25-word Recall	0.16	0.11	0.16	0.64	-0.12
List 1 recall	0.19	0.12	0.16	0.64	0.30
List 2 recall	0.17	0.12	0.16	0.75	0.50
Fruit words	0.26	0.06	0.27	0.40	1.09
Furniture words	0.20	0.05	0.20	0.31	0.02
M words	0.30	0.07	0.30	0.24	0.09
S words	0.33	0.08	0.32	0.09	-0.29
Recollect ROC	0.24	0.15	0.14	0.64	-0.17
Famil ROC	0.66	0.41	0.37	-0.15	0.59

Note. Recollect = Recollection, Famil = Familiarity, Cattell = Cattell Culture Fair Matrices, RAPM = Raven's Advanced Progressive Matrices.

The descriptive statistics for each of the tasks and the recognition memory model parameters are presented in Table 4.2. As can be seen, there were no severe univariate violations of skew (|skew index| > 3.0) or kurtosis (|kurtosis index| > 10.0) and inspection of univariate distributions did not identify any additional outliers that were not already addressed. Consistent with the low hit rate on recognition tasks, the number of words recalled was low overall, with a mean between 4 and 5 words for each list. For the fluid intelligence tasks, participants on average successfully answered around 7 of the 16 problems for the RAPM, 8 of 20 problems for the Cattell Culture Fair matrices, and 10 out of 15 of the Number Series problems. The mean proportion correct on test trials for all of the PS tasks was .95 or greater, so mean RTs were not greatly influenced by varying number of correct responses. For the fluency tasks, participants averaged approximately 12 words for the category tasks and 18 words in the letter tasks during the 60 second time frame.

Table 4.3 Correlations with model parameters and cognitive measures

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Correct HR															
2. Recollect R/K	.64														
3. Famil R/K	.82	.15													
4. *mean-of-R*	.62	.55	.42												
5. *mean-of-F*	.19	*.06*	.20	-.59											
6. *SD-of-R*	.55	.65	.19	.36	.25										
7. *d′* UVSD	.94	.65	.76	.63	.22	.60									
8. V_{old} UVSD	.52	.42	.33	.32	.22	.69	.66								
9. Recollect ROC	.73	.62	.52	.48	.22	.52	.80	.65							
10. Famil ROC	.79	.42	.75	.51	.13	.36	.74	.18	.29						
11. RT dot	*.03*	*.07*	*-.01*	*.02*	*.01*	*-.00*	*.03*	*-.05*	*.03*	*.05*					
12. RT animal	-.18	*-.07*	-.18	*-.06*	*-.08*	*-.10*	-.17	*-.09*	*-.10*	-.15	.42				
13. RT shape	*-.05*	*-.02*	*-.04*	*.04*	*-.09*	*-.05*	*-.06*	*-.10*	*-.02*	*.01*	.57	.47			
14. RAPM	.32	*.13*	.31	.27	*-.02*	.22	.30	*.07*	*.11*	.36	-.21	*-.09*	-.16		
15. Number Ser	.27	.16	.18	.17	*.06*	.24	.23	*.04*	*.13*	.25	*-.14*	-.16	-.15	.53	
16. Cattell	.30	.22	.24	.21	*.05*	.21	.27	*.06*	.14	.28	*-.09*	-.08	*-.04*	.46	.39

Note. Non-significant correlations ($p > .05$) are italicized. All correlations with absolute values greater than or equal to .15 were significant, $p < .05$. All absolute values greater than or equal to .19 were significant, $p < .01$. Recollect = Recollection, Famil = Familiarity, Cattell = Cattell Culture Fair Matrices, RAPM = Raven's Advanced Progressive Matrices. R/K parameters are DPSD parameters based on the remember and know rates, ROC parameters are based on the ROC fit of the DPSD model.

4.2.1 Relationship with PS and gF

The correlations between the fluid intelligence, processing speed tasks, and recognition memory variables are presented in Table 4.3. As expected, the three gF tasks and the three PS tasks each exhibited at least a moderate and significant correlation with the other tasks within their construct. Also of note, the *mean-of-F* parameter from the CDP model did not exhibit a significant correlation with any of the non-recognition tasks. This was unsurprising due to the previously established lack of reliability of this parameter, and as planned, the parameter was excluded from all of the SEM models.

For each of the below analyses, data were evaluated for multivariate normality. For several of the analyses, there were minor deviations from multivariate normality identified. Specifically, the Mardia's skewness test was significant (p <.01) and evaluation of the plot of Malahanobis' distance confirmed slight skew to the data but did not identify extreme multivariate outliers. The skewness of data existed regardless of which of the recognition parameters were included in the analysis. To compensate for any potential issues that may arise from the violations of multivariate normality, robust estimates of model fit metrics and standard errors of estimates are reported for all models below.

Baseline Model

The first model assessed was a baseline model that did not include any of the recognition memory models' parameters. This model established the fit of the measurement model of PS and gF and the relationship between PS and gF. Corrected Hit Rate (Hit rate- False alarm rate) was included as a variable to evaluate the overall relationship of performance on the recognition memory task with the other constructs.

Measurement Model. Based on the results of Study 1, the estimated reliability value of .80 was used for Corrected HR. An estimated error variance of .005 was calculated for Corrected HR based on this reliability estimate. The baseline measurement model provided an acceptable fit to the data, χ^2 (12) = 18.75, p = .095. CFI = .98. The covariance with PS and Corrected HR was not significant, all other covariances were significant. Next, the structural regression model was fit to the data.

Structural Regression. The initial model fit was good, but the path from PS to Corrected HR was non-significant. This path was removed from the model and the revised model did not significantly change in fit as determined by a non-significant chi-squared and a change in CFI of less than .01, χ^2(1)= .629, p = .427, ΔCFI =.001). The revised model fit the data well, χ^2 (13) = 18.65, p = .143, CFI = .98. Additional model fit statistics for this SEM model and the best-fitting models from the analysis of each recognition memory model with gF and PS are provided in Table 4.4. However, it should be noted that these fit indices are not directly comparable for the different recognition memory models because they are based on different variance-covariance matrices.

The structural components (path coefficients) of these models are reported in Table 4.5. As can be seen in Table 4.5, the path from Corrected HR to gF was significant, and the standardized coefficient was .47. The total variance explained (R^2) in gF by the combination of Corrected HR and PS was .28. Bootstrapping was used in order to create confidence intervals around the overall R^2 value for gF as well as the standardized coefficients leading to gF. Five thousand bootstrap samples were drawn, and the BCa confidence intervals around the R^2 estimate were 95% CI [.13, .46]. The 95% CI for the standardized coefficients are listed in Table 4.5. This model established that PS and gF were related in this young adult sample, and that recognition performance, as characterized by Corrected HR, was a valid predictor of gF. However, Corrected HR was not associated with PS.

Next, the parameters of the recognition memory models were compared. These analyses were designed to help determine which, if any, model offers best prediction of gF as wells as whether any specific parameters do the best job at capturing the relationship between gF and recognition. First, the results from the DPSD model parameters based on remember and know judgments are reported.

Table 4.4. Model fits for best-fitting Structural Equation Models with PS and gF.

Model	Chi-sq	df	p value	CFI	RMSEA	SRMR	AIC	BIC
Corrected HR (Baseline)	18.65	13	.143	.98	.05 [.00, .09]	.04	-1830.42	-1758.76
DPSD R/K	25.47	18	.113	.97	.05 [.00, .08]	.04	-2037.90	-1953.21
CDP 2-factor	13.41	18	.767	1.000	.00 [.00,.05]	.03	-595.08	-510.38
CDP *Mean-of-R*	15.39	19	.698	1.000	.00 [.00, .05]	.04	-601.83	-546.54
UVSD D-prime	17.31	13	.186	.985	.04 [.00, .09]	.04	-1248.15	-1176.49
DPSD ROC 2-factor	19.81	18	.343	.994	.02 [.00, .07]	.04	-1669.23	-1610.69
DPSD ROC F-only	23.49	19	.216	.985	.03 [.00,.07]	.05	-1667.71	-1612.42

Note. 90% Confidence Intervals in Brackets. CFI = Comparative Fit Index, RMSEA = Root Mean Square Approximate, SRMR = Standardized Mean Residual, BIC = Bayesian Information Criteria, AIC = Akaike's Information Criteria. Because models were not based on the same variance-covariance matrix, the fits of models are not directly comparable.

DPSD model from Remember-Know Judgments

Measurement Model. Based on the results of Study 1, an estimated reliability of .80 was used for both recollection and familiarity. Error variances of .005 for recollection and .004 for familiarity were calculated and used in model fitting. The measurement model allowed all latent variables to covary, except recollection and familiarity. Because recollection and familiarity were single indicator factors, the covariance between the two had to be entirely due to error or latent covariance. Recollection and familiarity are defined as independent processes, so the latent covariance between the two was set to zero and the error covariance was freely estimated. The degrees of freedom of the model and overall model fit were not affected by this choice. The measurement model was a good fit to the data, χ^2 (16) = 23.65, p = .100. CFI = .97. Neither familiarity nor recollection were significantly related to PS. All other latent covariances were significant.

Structural Regression. The structural regression model that included all hypothesized paths provided a good fit to the model. The model fit was not significantly changed when the non-significant paths between PS and recollection and familiarity were removed, χ^2(2) = 1.11, p = .575, ΔCFI =.001. Overall, as can be seen in Table 4.4, this simplified model fit the data well χ^2 (18) = 25.47, p = .113, CFI = .97. Both recollection and familiarity independently predicted gF, indicated by significant path loadings ($p < .01$). As can be seen in Table 4.5, the standardized path coefficients from recollection to gF was .26 and from familiarity to gF was .39. The total variance explained in gF from the model (including PS) was estimated to be .28 and the confidence intervals around the estimate were 95% CI [.15,.50]. Figure 4.3 plots the estimated R^2 of gF for each SEM model and shows that the variance explained for this model was roughly equivalent to the Corrected HR model.

CDP Model

Measurement Model. Based on the results of Study 1, the estimated reliability value for both the *mean-of-R* and *SD-of-R* was .60. This translated to estimated error variances of .732 for *mean-of-R* and .228 for *SD-of-R*. The measurement model allowed all latent variables to covary. Because they are parameters of the same distribution it was expected that *mean-of-R* and *SD-of-R* were related, so their covariance was estimated at the latent level, as opposed to set as error covariance. The measurement model was an excellent fit to the data, χ^2 (16) = 12.59, *p* = .703. CFI = 1.00. Neither *mean-of-R* nor *SD-of-R* were related to PS. All other covariances were significant (*p* < .01).

Table 4.5. Path coefficients to gF

Parameter	Est	SE	z	*p*-value	Standardized est	R-sq
Baseline model						
Corrected HR	0.47	0.10	4.65	.000	0.47 [.27,.62]	.22
PS	-0.40	0.14	-2.94	.003	-0.24 [-.40, -.08]	.06
DPSD from R/K						
Recollection	0.27	0.09	2.84	.005	0.25 [.07, .41]	.06
Familiarity	0.47	0.12	3.83	.000	0.40 [.20, .56]	.16
CDP						
Mean-of-R	0.04	0.02	1.91	**.056**	0.28 [-05, .58]	.08
SD-of-R	0.06	0.03	1.72	**.086**	0.22 [-.12,.52]	.05
CDP single parameter						
Mean-of-R	0.06	0.02	4.01	.000	0.41 [.22,.58]	.17
UVSD						
d'	0.09	0.02	4.01	.000	0.41 [.22, .57]	.17
DPSD from Confidence						
Familiarity	0.25	0.05	5.22	.000	0.53 [.36,.69]	.29
Recollection	0.21	0.11	1.91	**.056**	0.18 [.00, .35]	.03
DPSD single parameter						
Familiarity	0.24	0.05	5.05	.000	0.50 [.32, .66]	.25

Note. The PS path to gF was significant and roughly equivalent in all models and is only shown in the baseline model. The path to PS and each of the recognition parameters was not significant. Values in brackets are bias-corrected and-accelerated 95% confidence intervals based on 5,000 bootstrap samples.

Structural Regression. The structural regression model that included all hypothesized paths provided a good fit to the data, χ^2 (18) = 13.41, *p* =.767, CFI = 1.00. The model fit was not decremented by removing the non-significant paths between PS and both of the recognition memory parameters, χ^2 (2) = .84, p = .660, ΔCFI =.000. The model fit for this final structural regression model is provided in Table 4.4. Overall, the model fit the data well, however, as can be seen in Table 4.5, neither the *SD-of-R* parameter's path to gF (p = .094) nor *mean-of-R* to gF (p = .056) was significant. The total variance explained in gF from the model (including PS) was estimated to be .28 and the bootstrapped confidence intervals around the estimate were [.15,.48].

The *SD-of-R* and *mean-of-R* were highly correlated parameters (latent r = .61), and thus did not appear to have independent effects. An additional model with the path from *SD-of-R* to gF set to zero was estimated. This model had an excellent fit to the data, χ^2 (19) = 15.38 p = .698, CFI = 1.00 and the overall model fit was not significantly worse, χ^2 (1) = 2.46, p = .169, ΔCFI =.00. The standardized path coefficient from *mean-of-R* to gF increased to .42 and overall, the explained variance was .25, 95% CI [.13, .46], similar to the model that included a path from *SD-of-R*.

UVSD Model

Based on the results of Study 1, the estimated reliability of d' was .80 and .55 for V_{old}. This translated to estimated error variances of .106 for d' and .053 for V_{old}. The measurement model allowed all latent variables to covary. Because it was expected that discriminability (d') and variance of the old distribution would be theoretically related, their covariance was estimated at the latent level as opposed to set as error variance. However, the two variables were too highly related to fit the model, as the estimated correlation between the two was greater than 1.0. When instead the shared variance was set to be entirely due to error covariance, this was also greater than 1.0 in the standardized models. So, the two parameters were not independent in this sample and could not be included as separate factors in the same model. Instead, a measurement model with just *d′* was estimated, consistent with the UVSD argument that a single memory parameter explains performance on recognition. This model was a good fit to the data, χ^2 (12) = 16.45, *p* = .170, CFI = .98.

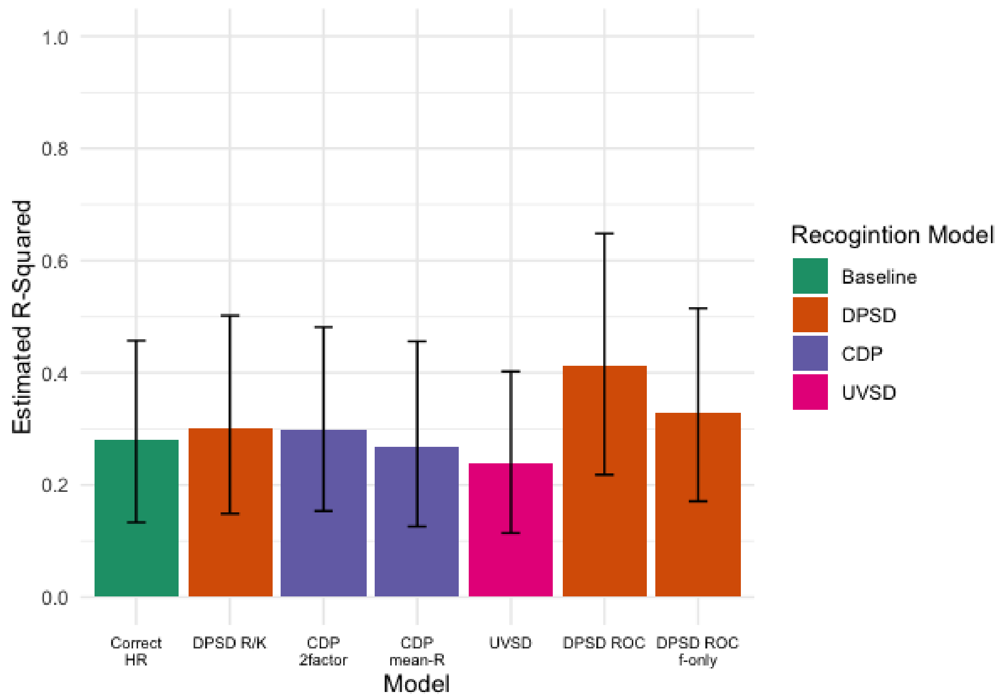


Figure 4.3. Estimated variance explained (R^2) in gF by combination of PS and recognition memory parameters. Error bars are BCa estimates of 95% confidence intervals based on 5,000 bootstrap samples.

Structural Regression. The structural regression model that included all hypothesized paths provided a good fit to the data, χ^2 (12) = 15.97, p =.193, CFI = .99. The model fit was unchanged by removing the non-significant paths between PS and *d′*, χ^2 = .782, ΔCFI =.001. The fit for this model is provided in Table 4.4. Overall, the path coefficient from *d′* to gF was significant (p <.01) and the standardized path coefficient was .41. The R^2 for gF for the model was estimated to be .23 and the bootstrapped confidence intervals around the estimate were 95% CI [.11,.40].

DPSD Model based on ROC

The final model assessed was the DPSD based on ROC data. Based on the results of Study 1, the estimated reliability of the ROC estimate of recollection was .75 and of familiarity was .65. This translated to estimated error variances of .006 for recollection and .060 for familiarity. The measurement model allowed all latent variables to covary expect recollection and familiarity. As in the R/K DPSD model, the error covariance between recollection and familiarity was freely estimated. When the error covariance was estimated, the correlation between recollection and familiarity was extremely high (.92), although below 1.0, allowing the model to be estimated. This model was a good fit to the data, χ^2 (12) = 16.45, p = .17, CFI = .984. Neither recognition memory parameter was related to PS, but all other covariances were significant (p < .05)

Structural Regression. The structural regression model that included all hypothesized paths provided a good fit to the data. As with the other models, the model fit was unchanged by removing the non-significant paths between PS and recollection and familiarity, χ^2 (2) = .08, p = .960, ΔCFI = .007. Overall, the resulting model fit the data well, χ^2 (18) = 19.81, p = .343, CFI = .99. As can be seen in Table 4.5, the standardized path coefficient from recollection to gF was .18 and from Familiarity to gF was .54. However, the path from recollection to gF was not significant, albeit barely so (z = 1.91, p = .056). The total variance explained in gF from the model (including PS) was estimated to be .38, 95% CI [.22, .65]. This R^2 was the highest of any model. However, due to the non-significant path from recollection, a model which constrained this coefficient to be equal to zero was also tested. The model fit was not significantly worse, χ^2 (1) = 3.64, p = .056, ΔCFI = .009. The model with only familiarity and PS predicting gF accounted for 32% of the variance in gF, 95% CI [.17, .51], which was similar to the proportion of variance explained by the models from the other recognition memory models.

4.2.2 Relationships with Recall

Next, the association between recall performance and parameters from each of the recognition memory models was assessed. The correlations of the model parameters including the recall variables can be found in Table 4.6. All correlations were significant ($p < .01$) between performance on each of the recall tasks and every recognition memory parameter (absent *mean-of-F*, which was excluded from the analysis). The correlations were generally higher in magnitude between the recall tasks that occurred after List 1 and 2 of the recognition task than for the standalone recall list of 25 words. This was unsurprising given that the recognition parameters were based on recognition of those two lists of words.

Table 4.6 Correlations of recognition model parameters with recall

Variable	1	2	3
1. 25-word recall			
2. R/K recall 1	.42**		
3. R/K recall 2	.29**	.44**	
4. Correct HR	.30**	.52**	.51**
5. Recollect R/K	.25**	.39**	.40**
6. Famil R/K	.24**	.41**	.37**
7. *mean-of-R* CDP	.28**	.41**	.34**
8. *SD-of-R* CDP	.25**	.32**	.34**
9. d' UVSD	.33**	.52**	.55**
10. V_{old} UVSD	.23**	.27**	.36**
11. Recollect ROC	.25**	.44**	.53**
12. Famil ROC	.27**	.44**	.33**

Note. Full correlation matrix of recognition parameters can be found in Table 4.3. Recollect = Recollection, Famil = Familiarity. ** indicates $p < .01$.

DPSD from R/K

First, a single-factor model was fit including the three recall tasks and the R/K-based recollection parameter loading on one latent factor. This model fit the data well, χ^2 (2), 3.175, *p*

= .204, CFI = .99. Additional model fit indices of this model and other model fits of recall models are provided in Table 4.7. The recollection parameter loaded significantly on the single factor. The standardized loading of the recollection parameter was .56 and all of the coefficients for can be found in Table 4.8. The composite reliability of the factor was decent, .70, and the AVE for the indicators was .37, less than .50, which has been suggested as a threshold for a good model (Kline, 2016). A separate model was conducted to assess the relationship with familiarity and recall. As in the analysis of gF, a single-indicator familiarity latent variable was included in the model. The fit of this model was good, χ^2 (3), 4.26, p = .37, CFI = 1.00. The familiarity factor was highly correlated with the recall factor (r = .62).

Table 4.7. Model fit indices for CFA models with recall

Model	Chisq	df	p value	CFI	RMSEA	SRMR	AIC	BIC
DPSD R/K	3.17	2	.204	.99	.06 [.00, .17]	.03	-1101.88	-1075.82
CDP *mean-of-R*	0.37	2	.830	1.00	.05 [.00, .09]	.01	-276.69	-250.63
CDP *SD-of-R*	2.13	2	.345	1.00	.02 [.00,.15]	.02	-488.69	-462.63
UVSD V_{old}	5.40	2	.067	.97	.09 [.00, .20]	.04	-786.98	-760.93
UVSD *d′*	6.92	2	.031	.97	.11 [.01, .15]	.03	-563.92	-537.86
UVSD *d′* (with error)	.09	1	.765	1.00	.00 [.00,.13]	.00	-568.29	-538.97
DPSD ROC	13.32	2	.001	.94	.17 [.08,.28]	.05	-1143.39	-1117.33
DPSD ROC (with error)	0.48	1	.487	1.00	.00 [.00, .00]	.01	-1147.40	-1105.05

Note. UVSD *d′* with error is model with error covariance estimated between *d′* and recall at time 2. DPSD ROC with error is model with error covariance estimated between recollection and recall at time 2. Brackets indicate 90% confidence intervals.

CDP

First, the model was fit with just the *mean-of-R* parameter loading on the recall factor. This model fit the data well, χ^2 (2), .373, p = .830, CFI = 1.00. The loading of the *mean-of-R* parameter was .54. and the factor composite reliability was .41, suggesting poor reliability, with an AVE = .29. A separate model tested whether the *SD-of-R* parameter loaded on the recall factor. A single-factor model with four manifest variables included in the model, also fit the data well, χ^2 (2), 2.13, p = .345, CFI = 1.00. The *SD-of-R* variable significantly loaded on the factor and the standardized coefficient was .48, suggesting a relatively weak indicator. For the *SD-of-R* model the composite reliability was .42, also suggesting poor reliability, and the AVE was .24.

Table 4.8 Variable loadings on Recall factor

Parameter	Est	SE	z	*p*	CI lower	CI upper	Stand	R-sq
<u>DPSD R/K</u>								
Recall 25-word	1.00	0.00	NA	NA	1.00	1.00	0.52	.27
RK1 recall	1.51	0.26	5.80	.000	1.00	2.02	0.74	.54
RK2 recall	1.25	0.30	4.22	.000	0.67	1.83	0.62	.39
Recollection	1.48	0.35	4.21	.000	0.79	2.17	0.56	.31
<u>CDP model</u>								
Recall 25-word	1.00	0.00	NA	NA	1.00	1.00	0.54	.29
RK1 recall	1.54	0.27	5.646	0.000	1.01	2.08	0.77	.59
RK2 recall	1.13	0.23	4.93	.000	0.68	1.58	0.58	.34
Mean-of-R	11.98	2.28	5.27	.000	7.52	16.44	0.54	.29
<u>UVSD model</u>								
Recall 25-word	1.00	0.00	NA	NA	1.00	1.00	0.52	.26
RK1 recall	1.68	0.35	4.81	.000	0.99	2.36	0.81	.66
RK2 recall	1.10	0.22	5.13	.000	0.68	1.52	0.55	.30
d′	7.87	1.42	5.54	.000	5.08	10.65	0.64	.41
<u>DPSD ROC</u>								
Recall 25-word	1.00	0.00	NA	NA	1.00	1.00	0.51	.26
RK1 recall	1.73	0.40	4.34	.000	0.95	2.51	0.83	.68
RK2 recall	1.11	0.22	5.10	.000	0.68	1.53	0.54	.29
Recollection	1.35	0.29	4.72	.000	0.79	1.91	0.52	.27

Note. Coefficients presented in the table are for best fitting CFA model for each recognition memory model. The characteristics of additional models are described in the text.

UVSD

The two key parameters from the UVSD model were separately evaluated for their relationship with recall. First, the model was fit with just the V_{old} parameter loading on the recall factor. This model fit the data well, χ^2 (2), 5.45, p = .067, CFI = .97. The loading of the variance parameter was .45, suggesting a relatively weak relationship with the factor. The composite reliability was .51 and the AVE was .25. Next, an additional model was fit with only the d' parameter. However, this model was not a close fit to the data, χ^2 (2), 6.92, p = .031, CFI = .97, according to the χ^2 test. The model fit indices indicated mixed evidence for overall fit, with the CFI and SRMR within accepted ranges, whereas the RMSEA did not indicate a close fit (Table 4.7). The d' parameter exhibited a strong loading structural coefficient on the single-factor (.76). The composite reliability was .71 and the AVE was .57, which was greater than the suggested threshold of .50.

Evaluation of the modification indices suggested a closer-fitting model would allow the errors from R/K recall List 2 and d' to correlate. This modification made theoretical sense because List 2, along with List 1, was the study list used for the recognition task that estimation of the d' parameter was based on. This model fit the data significantly better, χ^2 (1), 7.58, p = .006, ΔCFI = .026. The standardized loading of d' for this model was .64. This reduced the factor's composite reliability to .56 and the AVE to .41.

DPSD from ROC

First, the model was fit with just the recollection parameter loading on the recall factor. This model did not fit the data closely, χ^2 (2), 13.32, p = .001, CFI = .94. The loading of the recollection parameter was .68. For this factor, the composite reliability was .73 and the AVE was .42. As in the UVSD analysis, evaluation of the modification indices suggested allowing the errors from R/K recall List 2 and recollection to correlate would lead to a closer-fitting model. This model fit the data significantly better, χ^2 (1), 16.73, $p < .001$, ΔCFI = .057. The standardized loading of *d′* for this model was .64. However, this reduced the composite reliability to .65 and the AVE = .37.

When a separate familiarity variable was included in the model, this factor was highly correlated with the recall factor ($r = .68$). This model was also a close fit, χ^2(3), 1.00, $p = .006$, CFI = 1.00.

Combined Model Analysis

Although the above analyses provided an overview of the fit of each of the model's key parameters with recall, they did not provide a direct comparison between the models. To compare model parameters, all the models' strongest indicators were assessed in the same model. For this analysis, only one estimation of the DPSD was used in order to limit the number of parameters included in the model. The R/K version of DPSD was used because this estimation was *a priori* chosen to be the primary estimate of the DPSD model. The above analysis also suggested that estimation of recollection had a close fit with recall. The parameters included in the model were therefore, *d′*, *mean-of-R*, and the R/K estimate of recollection, which were all z-scored, along with the three recall scores. The error variances between the three recognition memory parameters were freely estimated because they were calculated from the same recognition memory task. The overall fit of the model was first assessed with the constraint that each of the recognition memory parameters pattern coefficients (loadings) were equal. This model was a relatively poor fit to the data, χ^2 (8) = 19.32, $p = .013$, CFI =.97. The standardized coefficient for each of the recognition memory parameters was .65.

The next step was to free one of these equality constraints. First, the constraint that the loading of DPSD recollection was equal to the other two recognition memory parameters was

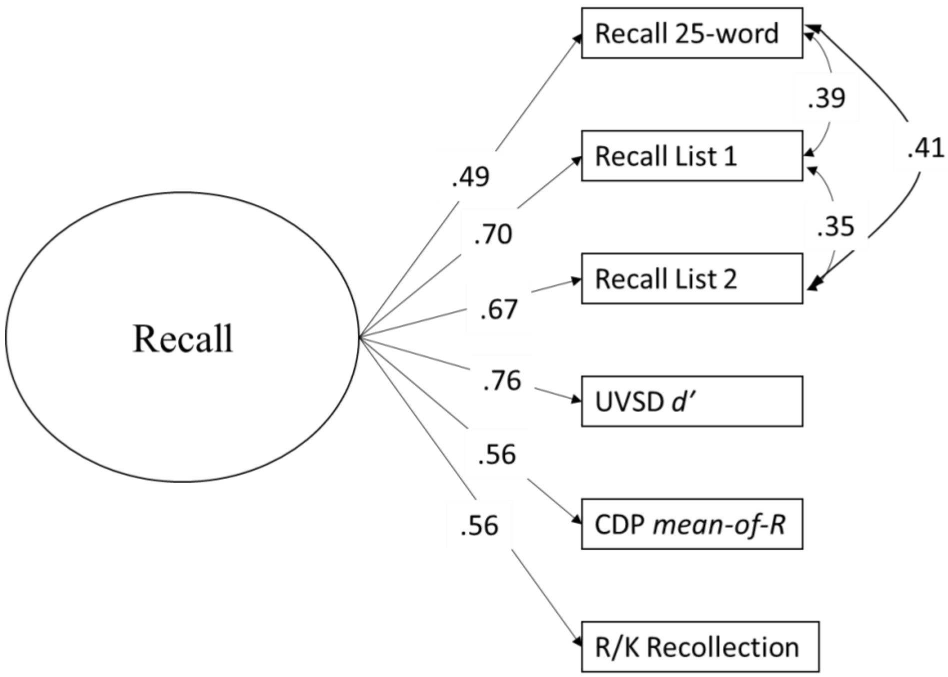


Figure 4.4. Recall model with parameter from each model. Arrows from circular to rectangle indicate standardized (structural) coefficient. Curved arrows equal correlated errors. All loadings and error covariances were significant (*p* <. 05). In this model, coefficients for DPSD recollection and CDP mean-of-R were set to be equal.

lifted. Freeing this constraint did not provide a significantly better fit to the data, $\chi^2(1) = 3.50$, $p = .061$, ΔCFI =.006. The standardized coefficient for recollection was .56 and for each of the recognition memory parameters was .68. Then, a model in which the equality of the *mean-of-R* parameter with the other two parameters was eliminated was assessed. The standardized coefficient for *mean-of-R* was .55 and for each of the recognition memory parameters was .69. Freeing this constraint did not provide a better fit to the data, $\chi^2(1) = 3.03$, $p = .081$, ΔCFI =.007. Finally, the constraint that *d′* was equal to the other parameters was lifted. Freeing this constraint did lead to a significantly closer-fitting model, $\chi^2(1) = 11.06$, $p < .001$, ΔCFI =.26. Figure 4.4 illustrates this model and demonstrates that the *d′* coefficient (.76) was larger than that of the two versions of recollection (.56). Because this model was the best fitting model, *d′* was determined to be a comparatively stronger indicator of this factor.

4.2.3 Relationships with Fluency

Initial fitting of a measurement model indicated that the verbal fluency construct was not significantly related to the other constructs (PS, gF, and recall). As a result, this construct was not included in the main analysis but was evaluated separately. Models were fit in which parameters from each recognition memory model were assessed as single-factor latent variables and the correlation was obtained with a fluency latent variable. The parameters assessed were familiarity from both DPSD estimates, *d′*, and *SD-of-R*. The correlation matrix for this model can be found in Appendix B and suggested weak correlations (r = < .20) between each of the fluency tasks and each of the recognition parameters. In each model assessed, the covariance between the latent variables did not approach significance and the latent correlation was estimated to be less than r = .10. As a result, no further tests of models with fluency were conducted.

4.2.4 Multiple Regression

In addition to the SEM analyses, multiple linear regression was used to test if the predictors of the models significantly predicted three variables; gF, PS, and recall ability. These analyses were designed to help confirm that the conclusions made from the SEM analyses were not due to any assumptions (e.g., about reliability) made in the SEM models as well as to test models that could not be fit due to estimated latent correlations greater than 1.0.. The outcome variables were the predicted scores on the three latent variables for the participants, which were z-scored so that their scale was in standard deviation units. . The analysis of PS, as in the above SEM models, found that none of the key parameters from any of the recognition decision models were significant predictors of PS, thus those results are not reported further.

gF

For the R/K version of the DPSD, recollection and familiarity were allowed to predict gF. The overall regression was statistically significant, $R^2 = .16$, $F(2, 189) = 18.31$, $p < .001$. The complete regression equation table for all the gF models is presented in Appendix C. Both the recollection (ß = 0.19, $p = .006$) and familiarity (ß = 0.33, $p < .001$) parameters significantly predicted gF.

For the CDP model, mean of recollection and familiarity and *SD-of-R* were allowed to predict gF. The overall regression was statistically significant, $R^2 = .18$, $F(3, 188) = 13.38$, $p < .001$. Both the *mean-of-R* (ß = 0.46, $p < .001$) and *mean-of-F* (ß = 0.29, $p = .007$) parameters significantly predicted gF. However, the *SD-of-R* parameter did not significantly predict gF (ß = 0.07, p = .447).

For the UVSD, d' and V_{old} were allowed to predict gF. The overall regression was statistically significant, $R^2 = .18$, $F(2, 189) = 21.34$, $p < .001$. Both the d' (ß = 0.55, $p < .001$) and V_{old} (ß = -0.25, $p = .006$) parameters significantly predicted gF. However, the direction of the V_{old} parameter was negative. This is likely due to multicollinearity based on the high correlation between the two predictors from the model. This same issue led to the estimated latent correlation between the two being greater than 1.00 in the SEM model.

For the ROC version of the DPSD, recollection and familiarity were allowed to predict gF. The overall regression was statistically significant, $R^2 = .18$, $F(2, 189) = 20.53$, $p < .001$. The familiarity (ß = 0.39, $p < .001$) parameter significantly predicted gF. However, the recollection (ß = 0.09, $p = .180$) parameter did not significantly predict gF.

Overall, the results of these analyses converged with the results from the SEM analyses, by finding that the models did not vary greatly in their ability to predict gF (all R^2 between .16

and .18). The *mean-of-F* parameter was surprisingly a significant predictor of gF; however, this did not lead to a higher R^2 for the CDP model than the other models, which had two predictors compared to three for the CDP.

Recall Ability

For the R/K version of the DPSD, recollection and familiarity predicted recall ability. The overall regression was statistically significant, $R^2 = .36$, $F(2, 189) = 53.19$, $p < .001$. The complete regression equation for all the gF models is presented in Appendix C. Both the recollection (ß = 0.38, p = <.001) and familiarity (ß = 0.31, $p < .001$) parameters significantly predicted recall.

For the CDP model, mean of recollection and familiarity and *SD-of-R* were allowed to predict recall. The overall regression was statistically significant, $R^2 = .39$, $F(3, 188) = 39.63$, $p < .001$. Both the *mean-of-R* (ß = 0.78, $p < .001$) and *mean-of-F* (ß = 0.54, $p = .007$) parameters significantly predicted recall. However, the *SD-of-R* parameter was not (ß = -0.02, p = .813) a significant predictor.

For the UVSD, d' and V_{old} predicted recall ability. The overall regression was statistically significant, $R^2 = .38$, $F(2, 189) = 57.33$, $p < .001$. The d' (ß = 0.68, $p < .001$) and V_{old} parameter significantly predicted recall ability. However, the V_{old} parameter (ß = -0.11, $p = .135$) was not a significant predictor.

For the ROC version of the DPSD, recollection and familiarity predicted recall ability. The overall regression was statistically significant, $R^2 = .38$, $F(2, 189) = 57.52$, $p < .001$, and the complete regression equation for all the gF models is presented in Table 4.10. Both the recollection (ß = 0.39, $p < .001$) and familiarity (ß = 0.38, $p < .001$) parameters significantly predicted gF.

4.3 Discussion

The study was designed to evaluate the DPSD, the CDP, and the UVSD by assessing their parameters relationship with established cognitive constructs, including gF, PS, and recall. Assessing these relationships had three purposes; to assess evidence for convergent validity of the models, to provide better descriptions of each model's key parameters, and to use these to compare the models. These purposes build upon each as a better understanding of the models' parameters and convergent validity can help determine the one that is most useful in assessment of individual differences in recognition memory.

4.3.1. Convergent Validity

The results of the analysis of the relationship of model parameters with gF, as defined by the latent variable consisting of Raven's progressive matrices, Cattell's Culture Fair matrices, and a Number Series task. Each of the models had at least one parameter that significantly predicted gF, suggesting that each model exhibited some convergent validity. Of the models, only the R/K-based DPSD estimations had two parameters that significantly predicted gF.

As in the analyses with gF, the analyses of relationships with recall determined that each model had one parameter that loaded relatively strongly on a recall factor. Therefore, each model exhibited at least some evidence of convergent validity with recall as well. Except for the UVSD model, these models were a good fit to the data. The UVSD model was a better fit to the data when the error variance for d' was allowed to correlate with the recall score for Part B, a modification that made theoretical sense due to the scores being from partially the same study list. Kline (2016) has suggested that the average variance explained (AVE) by a CFA model should be .50 (50%) and in an ideal model the majority of variance (>.50) should be explained in each indicator. None of the recall models met either of these criteria, with the DPSD from R/K model exhibiting the highest AVE and reliability. Any indicator with an R^2 less than .50 is a candidate to be removed from a model, however, a more liberal criterion is that an acceptable indicator should have at minimum an R^2 of at least .30, corresponding to a standardized pattern coefficient of approximately .55 for an observed variable that only loads on one factor, in order to be included in a model. Slightly above 30% of DPSD recollection from the R/K model and *mean-of-R* from the CDP were explained and slightly below (.27) was explained for recollection estimated from the ROC procedure. Each of these parameters can be seen as exhibiting acceptable, but less than ideal, convergence with recall.

As noted in the discussion of Study 1, how related the parameters from the two estimations of DPSD parameters (R/K and ROC) can help establish the overall validity of the DPSD. The correlation between the two versions of recollection was .62, 95% CI [.53, .70]. The correlation between two versions of familiarity was .75, 95% CI [.68, .81]. These correlations suggest relatively high agreement between the two estimation procedures.

In general, the domain-general PS factor was not very related to recognition memory in younger adults. This was true of Corrected Hit-Rate, a performance metric that was not directly tied to any one model, as well as parameters that were hypothesized to be related to PS, such as familiarity and d'. In addition, verbal fluency was not significantly related to any other constructs nor any of the recognition memory parameters.

4.3.2 Model Description

The results of the analysis of the relationship of model parameters with gF identified the signal-detection based parameters, such as familiarity and d', as key predictors. Across models, the variables that explained the most variance (i.e., had the highest standardized path coefficient) in gF were familiarity (both estimates) for DPSD, *mean-of-R* for the CDP, and d' for the UVSD.

Of those parameters, each was calculated as discriminability based on signal detection curves, except for the R/K-based familiarity, which has been shown to converge with estimates from signal detection curves (e.g., Koen & Yonelinas, 2016). Of these variables, only the *mean-of-R* parameter was purported to involved recollection, but it may have been more akin to familiarity as traditionally defined in the DPSD, because it also incorporated confidence decisions and the CDP estimates did not produce a reliable familiarity parameter.

Based on the descriptions of dual-process models, recollection was predicted to be highly related to recall whereas familiarity was predicted to be only modestly related at most. However, there was a relatively large (r = .63, .68 for the R/K and ROC estimates, respectively) relationship between familiarity and recall. Other studies have shown a relationship between familiarity and recall through both 'know' judgments (Mickes et al., 2012) and the process-dissociation procedure (McCabe et al., 2011). However, the relationship was larger than expected and suggests a lack of divergent validity for familiarity.

It is a somewhat surprising result that familiarity-like processes dominate the explained variance in gF and was also highly related to recall. Most previous studies have suggested that recollection that would be more predictive parameter of other constructs, such as Unsworth and Brewer (2009), which found that recollection predicted gF but familiarity did not. However, that study, and others evaluating individual differences in familiarity and recollection (Blankenship et al., 2015; Long et al., 2008) did not use direct parameter estimates of recollection and familiarity. Instead, those studies separated performance on tasks that were primarily based on recollection versus those that relied on both processes. Because recollection is hypothesized to be the major process in many LTM tasks, whereas there are comparatively few tasks that are purported to rely solely on familiarity, it is possible that previous studies have underestimated the role of a person's ability to make decisions based on familiarity. However, in this study there was a relatively low rate of recollection, resulting in mean recollection estimates of only .19 and .22 for the R/K and ROC respectively. It is possible that floor effects contributed to lack of variability in these parameters and in turn influenced the lower convergent validity of recollection. This same issue of low recollection could have influenced the convergent validity of the *mean-of-R* parameter from the CDP and the V_{old} parameter from the UVSD, which is related to recollection.

With this caveat of a potential influence of a lack of recollection-like decisions in this study, a key implication of this study is that the results were unable to establish convergent validity for the variance parameters estimated as part of the UVSD and the CDP. The V_{old} parameter from the UVSD could not be included as a separate predictor in a model predicting gF, because its variance overlapped with d' causing an estimated latent correlation greater than 1.0. When V_{old} was included as an indicator variable of recall, the loading on the latent factor was .45. This equals an R^2 of .20, which is much less than suggested at minimum for a good indicator variable (Kline, 2016). Similar conclusions can be made about the *SD-of-R* from the CDP. This parameter did not predict gF significantly when it was included in the model. In addition, *SD-of-R*'s standardized loading on the recall factor, .47, corresponds to an R^2 of .22, a similarly small amount of variance explained.

A key takeaway from the correlations between estimates between the various recognition parameters (Table 4.3) is that correlations between parameters from the different models were relatively high overall. UVSD's d' parameter appears to be almost uniformly related with overall performance not factoring in confidence, as captured by Corrected HR, $r = .94$, 95% CI [.92, .96]. Correlations were also high between d' and several other parameters from different models, unsurprisingly since it captured overall performance. In particular, the ROC estimates of the

DPSD showed strong relationships with d', both familiarity, r = .74, 95% CI [.66, .79] and Recollection, r = .80, 95% CI [75, .85]. These results show how despite the differences between models, more variance is shared between them than what differentiates them. This is also affects the ability for the analyses to identify differences between models.

Given the low rate of recollection, and thus greater role of familiarity in this study's recognition test, it is perhaps less surprising that familiarity was strongly related to recall. Although there does seem to be a strong relationship with familiarity and recall performance, it is important to make a distinction between a process and ability. Although recollection may be the process that is needed for recall, there is nothing specific about the DPSD model that suggests that an ability to make successful memory judgments based on familiarity would be unrelated to an ability to recall items.

4.3.3 Model Comparison

The primary comparison of the recognition memory models in this study was based on their parameters' ability to predict gF. According to the baseline model, the gF latent variable was significantly predicted by overall recognition memory performance, as characterized by Corrected HR. Therefore, it is unsurprising that each of the models had at least one predictor that significantly predicted gF. In terms of the standardized path coefficient, the familiarity parameter from the ROC estimate of the DPSD model was the greatest individual predictor from any of the models. However, the confidence that this estimate is significantly greater than any of the other parameters is relatively low, as evidenced by considerable overall of the bootstrapped of confidence intervals around the estimates of the standardized path coefficients for each model (Table 4.5). In addition, the overall R^2 of gF in each of the SEM models had considerable overlap of confidence intervals (Figure 4.3), suggesting that none of the models explained any more or less variance in gF. The same conclusion was reached when multiple linear regression was used instead of SEM.

Of the three models under comparison, the DPSD was the only model that had two parameters that independently predicted significant variance in gF. However, the relation was only significant for both parameters in the R/K based model. This suggests that this model has the potential to be the most useful model for individual differences research as it has two potential sources of explaining performance in an individual, which may be more or less predictive of various constructs. Both recollection and familiarity exhibited modest evidence for convergent validity with gF. However, in the ROC estimation of the DPSD model, whatever variance that the recollection parameter shared with gF was not significant (or sufficiently independent from that predicted by familiarity) and was a very small influence ($R^2 = .03$) even if the standard error was smaller. However, this parameter did exhibit a moderate relationship with recall, suggesting some evidence for convergent validity.

Although some evidence suggests that the DPSD outperformed the other models in predicting gF, the CDP's version of recollection, the *mean-of-R* parameter, was a better predictor of gF than recollection from the DPSD, regardless of which estimation (R/K or ROC) was used. *Mean-of-R* predicted 17% of the variance whereas, as mentioned, the recollection parameter from the confidence responses was not a significant predictor and the recollection parameter from the R/K responses only predicted 6% of the variance. Therefore, there is some evidence that including confidence in recollection may have some impact in improving its convergent validity. However, because the CDP model did not have a reliable familiarity parameter, another interpretation is that the *mean-of-R* parameter is capturing partially a familiarity process. All of the convergent validity of the model came from that one parameter, whereas the DPSD was able to separate the variance in performance into two sources.

When the best indicators of recall from each model were included in the same model, the UVSD d' parameter had the highest loading on the factor. By testing the effect of equality constraints, it was determined that the higher d' loading was critical to model fit. This result is evidence in favor a single-process model being the most useful model of individual differences in LTM. However, it should be noted that the single factor in this analysis may not have been a recall factor. CFA factors are given names by researchers, and while this factor could be called recall, the factor consisted equally of parameters from recognition tasks and recall tasks. In addition, performance on the standalone 25-word recall task had the lowest loading on this task. Therefore, although d' had the highest association with the latent factor, the factor may have included idiosyncratic qualities of those particular words, or a participants' attention during the study phase, as opposed to a latent recall ability *per se*. In other words, it is possible from this analysis alone that d' is the best at capturing a general recall ability. Nevertheless, this analysis suggests the possibility that d' captures recall ability better than recollection. This evidence needs to be weighed alongside the lack of unique convergent validity of the V_{old} parameter, which is potentially problematic for the model even if this parameter is not described by the UVSD theory as being a key memory ability.

Overall, the evidence comparing models was not definitive. This is perhaps not surprising given that the correlations between parameters from different models, as seen in Table 4.3, were rather high. The implications of the results from the two studies for model comparison will be further discussed in the next Chapter.

Chapter 5: General Discussion

The goal of this dissertation was to evaluate three recognition memory models for their reliability and convergent validity. The motivation was both to compare the models and to understand the models' key parameters better, under the guiding principle that no model is correct, but some are useful (Box, 1976). From this perspective, the findings provided some evidence as to which models may be useful and why. The overall results will be summarized by model, then the implications for comparison will be discussed, along with limitations and ideas for future directions.

5.1 Summary of Results by Model

5.1.1 DPSD

The DPSD suggests that there are two processes that impact recognition memory performance: recollection and familiarity. According to this model, successful recollection leads to high confidence responses and "remember" responses in the remember-know paradigm, whereas if recollection fails, decisions are based on familiarity, which is categorized by a signal detection process and "know" responses in the R/K paradigm. In the recognition memory paradigm used in this study, parameter estimates for the DPSD were calculated directly from the remember rates and the know rates for old and new items. Estimates were separately calculated from the ROC curve, which factored in the response probabilities for old and new items at each confidence threshold. Using maximum likelihood estimation, recollection and familiarity parameters were calculated, along with five confidence criteria. The results from Study 1 (Chapter 3) suggest that the parameters from this model are reliable regardless of the estimation procedure used. Overall, the R/K based estimates were slightly more reliable than the ROC based estimates, but each of the estimates exhibited Intraclass correlation coefficients (ICC) of both key parameters ranging from good/substantial (>.60) to excellent (> .80), by generally accepted rules of thumb (Cicchetti & Sparrow, 1981). In addition, the ROC estimates from the DPSD outperformed the other two models in predicting HRs.

The results from Study 2 (Chapter 4) indicated that the familiarity parameter predicted a large portion of variance in gF, regardless of the estimation method. In fact, the ROC estimate of familiarity had the highest standardized path coefficient (.23) to gF of any individual parameter. Surprisingly, both versions of recollection did not predict much variance in gF and only the R/K estimate of recollection was a significant predictor. The ROC recollection parameter was almost significantly greater than zero, but even if the marginally significant path was included, it only explained around 3% of the variance in gF.

The recollection parameter based on the R/K judgments loading on the recall factor was .57. Overall, this model created a recall factor with relatively good composite reliability (.70) and an AVE (.37) that was less than ideal, but higher than the other models. Given that recall is hypothesized to be based primarily on recollection, the recollection parameter loading on this variable is an important test of convergent validity, necessary to demonstrate that the recollection estimate was valid indicator of what it is purported to be measuring. Like in the analysis of gF, the ROC-based recollection estimate was a slightly less valid indicator of recall than the R/K-based estimate, indicated by the standardized loading on the factor being .52. (R^2 = .27) and weaker reliability of the model Contrary to the DPSD theory, the familiarity parameter was highly related to recall. This is not a novel finding, as other researchers have found that some recall responses were associated with “know” responses and other measures of familiarity (McCabe et al., 2011; Mickes et al., 2012). However, a strict interpretation of the DPSD would expect this relationship to be small to non-existent, whereas in this study the latent correlation between familiarity and recall was .62 for the R/K parameters and .70 for the ROC parameters.

Overall, the evidence for reliability for the DPSD was strong, and the evidence for convergent validity was also strong, with a couple exceptions. The first exception was that recollection was less predictive of gF than hypothesized. This was particularly true of the recollection parameters from the ROC estimate, which may have been affected by relatively low performance, as well as the nature of the task, in which the participant may have responded differently than they may have if asked to make confidence judgments without having to also make an R/K judgment. The other issue for the DPSD, was the independence of familiarity from recall was not met. This is a contradictory finding to descriptions of the DPSD and similar dual-process models, as the definition of recollection makes it the key process behind recall, whereas familiarity is defined as independent from recollection (and hence, recall). Findings of a small relationship between familiarity and recall is easy to reconcile with the DPSD as the two processes might be separate but, like most cognitive processes, not entirely independent. However, a high correlation between familiarity and recall, like that found here, suggests that recall is not based primarily on recollection, a key prediction of the DPSD, but instead more equally based on recollection and familiarity, at least in situations where recollection is relatively low.

An additional assessment of the DPSD was made by assessing the relation between estimates from the estimates from the R/K responses and confidence ratings. The correlation between the two estimation procedures in the 200-person sample was .62 for recollection and .75 for familiarity. These correlations suggest pretty good agreement, albeit less than equivalence. Advocates for the DPSD do not suggest that the two estimation procedures be identical, as each are only estimating the underlying process, and R/K judgments in particular rely on an individual's introspective awareness of the state of recollection (Yonelinas, 2002). Few researchers have estimated both for the same individuals; Koen and Yonelinas' (2016) reported correlations of .68 for recollection and .50 for familiarity, based on 30 participants. These correlations are similar to what was found here and suggest that overall, the agreement between the two estimation procedures was about as high as expected. The correlations suggest that the two estimation procedures offer similar but not identical estimates.

The familiarity parameter's large influence on gF was not predicted *a priori* but may offer some clarity as to what the familiarity parameter is measuring. This was a novel finding, as previous studies had not found a relationship between familiarity and gF (Unsworth & Brewer, 2009). Familiarity has been criticized as being vaguely and inconsistently defined (Hintzman, 2011), and although this result does not offer a new definition, it does suggest the ability to may be highly related to gF. Familiarity, as an ability, would be defined by the DPSD as the ability to *successfully* recognize old items when recollection fails. In both estimates of the DPSD, recollection and familiarity were positively correlated. So, individuals with high scores on the familiarity parameter are not relying on familiarity when recollection could serve. These individuals instead made more accurate memory decisions in the absence of recollection, which occurred for a large number of test trials for most participants in this paradigm, where recollection estimates suggested only around 20% of old items were recollected on average.

A participant with a strong familiarity "ability" could be considered a person with generally stronger memory. Signal detection theory is defined by making decisions in situations with considerable noise (Green & Swets, 1966). A stronger sensitivity would suggest better memory (i.e., a better signal) but also perhaps a better logical decision-making process. Nonrecollected words are characterized by some uncertainty and being able to logically assign confidence based on the signal and make decisions on whether this word was likely on the study list versus not, is likely key to higher familiarity estimates. The construct gF is defined as the ability to solve novel problems. A decision-making process in which participants must essentially solve the problem of whether they saw the item or not, in situations when the memory is not strong, may considerably overlap with the process of solving novel problems.

A second (non-exclusive) possibility is that greater general familiarity-type memories may be associated with better performance on gF tests more than greater recollection. Whereas a crystalized intelligence test may depend on being able to recollect information or fact, gF tests do not ask for these types of memories. Familiarity with the general strategies that have been successful in the past may be more beneficial to solving new problems than recollecting specific details about previous experiences solving the type of problems on tests of gF. Regardless of the exact reason behind the association found here, the finding suggests that researchers should further examine familiarity's potential importance as an individual difference variable.

5.1.2 CDP

The CDP, like the DPSD, proposes that recognition memory is based on recollection and familiarity, but differs in the definition of the processes. Both processes are defined by separate signal detection models. To fit the model to the data, the parameters included the mean (*mean-of-R*) and standard deviation of the target distribution (*SD-of-R*) for recollection and the mean of the target distribution for familiarity (*mean-of-F*; the standard deviation was fixed at 1.0), plus confidence criteria. Study 1 suggested that *mean-of-F* was not a reliable parameter, with the ICC (.23) well below the threshold for moderate reliability. The other key parameters, *mean-of-R* and *SD-of-R*, were moderately reliable but unlike both DPSD memory parameters, below the threshold for “strong” reliability. The lack of reliability of the *mean-of-F* would suggest that the model is over-parameterized. Too many parameters can lead to overfitting of a single dataset and lack of replication and generalizability. This is evidenced by the CDP’s parameters performing worse than the DPSD at predicting HRs for the opposite part of the study. However, the CDP did perform significantly better suggesting that the model may still have some benefit from its additional parameters. Nonetheless, the potential overfitting is an important shortcoming, as Babyak (2004) states, “ ‘findings’ that appear in an overfitted model don’t really exist in the population and hence will not replicate” (*p*. 411). The model had a total of 9 free parameters including confidence criteria. Although there are other possible explanations for the lack of reliability, both the DPSD (ROC-estimate) and the UVSD had only 7 total free parameters and showed stronger reliability in their parameters, suggesting that the number of free parameters is the simplest explanation.

Although the lack of reliability of one of its parameters argues strongly against the model as a useful model for individual differences research, the model was still further tested in order to see if any of its parameters had some utility. The results from Study 2 indicated that *mean-of-R* was strongly predictive of gF. The standardized coefficient (.41) exceeded that of the recollection parameter from DPSD, although their confidence intervals overlapped. However, given the familiarity parameter’s unreliability, all the model’s convergent validity was based on the recollection components. The *SD-of-R* component was not significantly predictive of gF and was only estimated to explain around 3% of the variance in gF. Therefore, despite the increased predictive validity of the *mean-of-R* compared to recollection from DPSD, this model predicted slightly less variance overall in gF, although again the estimates had substantial overlap of confidence intervals. Of the three potential parameters that were hypothesized to be related to an individual's cognitive ability, only one was demonstrated to be predictive of gF. Furthermore, although *mean-of-R* was more predictive of gF than recollection from the DPSD, it was not more predictive than familiarity, but rather had a smaller estimated standardized coefficient than either version of familiarity.

The results evaluating the parameters' relationship with recall led to similar results. Both the mean and *SD* of the recollection distribution were hypothesized to load significantly on the recall factor. The *mean-of-R* was a good indicator of recall, with a standardized loading of .55. but the standardized loading of *SD-of-R* (.47) was less than the threshold proposed for good indicators. Analysis that included both indicators in the mode suggested that there was no difference between the *mean-of-R* and the recollection parameter from the R/K-based DPSD, suggesting that no added predictive validity was given by the more complex calculation of recollection.

Overall, the evidence for the CDP was weak. The CDP is a more complex model than either the DPSD or the UVSD, as defined by requiring additional parameters, as well as requiring additional data collected than is normally collected in recognition tasks. Whereas the DPSD and UVSD were based only on either recognition ratings or remember-know judgments, both were factored into the CDP's parameter estimations. Because these parameters were based on more data than either of the other models, they ought to be contributing more towards predicting cognitive abilities that require similar process as recognition memory such as fluid intelligence and recall memory to justify this additional data collection. However, the results did not offer evidence of CDP parameters explaining more variance, but rather equal or less. In addition, the *mean-of-F* was not at all reliable within participants and the low correlations of this parameter with other constructs found in Study 2 confirm that this parameter did not capture anything meaningful about an individual's recognition memory abilities. Surprisingly. The *mean-of-F* parameter was a significant predictor of gF and recall in multiple regression models that included the other parameters from the model. This suggests that when accounting for the other model parameters, *mean-of-F* can be a meaningful parameter, but given its lack of reliability it is unclear how to interpret this finding and whether it would replicate.

The one potential strength of the CDP is that the *mean-of-R* parameter appeared to be a better predictor of fluid intelligence than recollection based on confidence ratings or remember-know judgments alone. It is therefore possible that a recollection parameter that factors in varying levels of confidence may be a better overall measure of recollection. However, since familiarity was entirely unreliable in the model, it is unclear that this model can simultaneously estimate both recollection and familiarity. Instead, *mean-of-R* may be capturing some combination of the processes closer to an overall memory ability rather than a separate recollection process.

5.1.3 UVSD

The unequal variance signal detection (UVSD) model suggests that a single memory strength signal detection process explains recognition memory performance, as opposed to separate contributions of recollection and familiarity. However, in addition to the general memory strength variable, d', the model posits that the variance of targets is unequal to, and specifically greater than, that of lures. These two parameters are the key parameters in the model, making it equally complex as the DPSD, in terms of number of parameters. However, the model is characterized as a single memory strength model, as it argues that d' is sufficient to explain all the memory processes underlying a recognition memory decision.

The results from Study 1 suggest that the d' parameter was strongly reliable (*ICC* = .80) and the variance parameter was moderately reliably (*ICC* = .54). The UVSD performed worse than either of the other two models in its predictions of HR and FAR, suggesting that there was a cost to the relatively lower reliability of the V_{old} parameter. The results from Study 2 suggest that the two parameters from the model were too highly related to independently predict gF. This multicollinearity affected both possible SEM models (leading to an estimated latent correlation greater than 1.0) and the more conventional multiple regression (leading to a negative coefficient for V_{old}). However, a model which only included d' was predictive of gF. This result is not necessarily inconsistent with the model's predictions, as the model does not argue that these two parameters are independent. In fact, with two highly correlated parameters the model essentially functions as a single process model, as it is often described. Despite some suggestion that the variance parameter functions similarly to recollection (Dede et al., 2013), it did not load highly on a recall factor (std coefficient = .42). The loading for d' was high (.72) although the overall fit of the model was not good. In addition, d' was shown to be a better predictor of the model than either the DPSD or CDP's recollection estimates.

Overall, the reliability and validity of the UVSD was supported, with the exception that the V_{old} parameter was not highly reliable nor showed any separate convergent validity. Both parameters were at least moderately reliable, and the d' parameter exhibited convergent validity with gF and recall. The fact that the variance parameter was not a useful parameter in terms of sharing variance with other cognitive abilities does lead to some question to the model's overall validity. V_{old} is perhaps designed to be a noisy variable that captures non-memory processes and idiosyncrasies of the study material. Furthermore, it exhibited moderate reliability and it is possible that it is related to cognitive abilities that were not assessed in this dissertation, such as attentional control. However, if it does not, then it is simply a model parameter without a behavioral or cognitive basis, perhaps undermining the utility of this model.

5.2 Model Comparison

Based on the results of both studies, the CDP model did not show evidence for being a useful model for individual differences research. Despite the CDP's additional complexity, it did not outperform the other models and one of its key parameters lacked reliability. However, the results indicated at least modest support for the DPSD and UVSD's reliability and convergent validity. So, which model is more useful? These two studies provided some evidence by which to compare the two models to help determine which is more useful, at least for evaluating individual's recognition memory abilities.

Much of the evidence from Study 2 suggested that one memory parameter can explain the majority of shared variance between recognition memory and other cognitive abilities. To recap, UVSD's d' parameter explained a similar amount of variance in gF as both recollection and familiarity combined. Even when both recollection and familiarity were considered together the majority of variance was explained by only one of these variables, familiarity. In the analysis of recall, d' was shown to have a higher association with a latent recall variable than recollection. All these findings suggest that a single memory strength variable such as d' is all that is needed to explaining individual differences in recognition and their relation to other cognitive abilities. In addition, as discussed, familiarity's large relation to recall was large suggests that there is a greater relation between memory process than is typically suggested by the DPSD and similar dual-process models. If all memory tasks are moderately to strongly related, as previously argued by Unsworth (2019), this favors a single-process model memory strength model, such as the UVSD.

Despite the evidence for a large amount of recognition performance's shared variance with other cognitive abilities being explained by a single variable, the DPSD's predictions were largely upheld. Recollection, as calculated from the remember-know judgments, was a significant predictor of gF and loaded on the recall factor. This made the DPSD the only model that had two parameters that exhibited convergent validity. On the other hand, the UVSD's V_{old} parameter did not relate significantly to gF and had a relatively low loading on recall. The UVSD model is often described as a single process model, but it has been repeatedly demonstrated that the unequal variance parameter is needed to fit ROC data (e.g., Glanzer et al., 1999). If it is necessary to calculate two parameters in order to fit the data, it seems reasonable that the model with two parameters that exhibit meaningful convergence with other cognitive variables is preferable to one in which only one of the two parameters exhibits convergent validity. In addition, in Study 1, when the parameters from the models were used to predict HR and FAR, the DPSD's predicted HR were closer to the observed HR than the UVSD.

If the single parameter, d', explained more variance in gF than either parameter from the DPSD, then it could be argued that the model is superior, even given that the variance parameter was not able to capture separate variance. However, the standardized path from the familiarity parameter from the DPSD from R/K judgments to gF (.40) was nearly equivalent to that of the UVSD (.41). In addition, the path from familiarity from the ROC data (.50), which uses the same data as the UVSD, was greater than that of d', although the confidence intervals substantially overlapped. So, the evidence suggests that familiarity captures as much meaningful variance as d', and that recollection provides a separate, albeit small, source of meaningful variance in performance. This suggests that there is no cost, in terms of reducing the predictive validity of familiarity, by separating out recollection, and there are potential benefits, although they were not definitively found here.

In this particular paradigm, recollection did not explain a lot of variability in gF nor did it predict recall as well as d', but this may change in different paradigms or different populations. Indeed, several studies have shown that recollection is related to other cognitive variables, whereas the role of familiarity as a predictor of cognitive abilities was in question (Blankenship et al., 2015; Long et al., 2005; Unsworth & Brewer, 2009). On the other hand, this study showed that familiarity scores played a surprisingly large role as a predictor of gF, therefore separate studies show evidence for convergent validity of both parameters. Of course, previous studies of individual differences in recollection and familiarity did not directly estimate the recollection and familiarity parameters from recognition, but rather inferred them from performance on specific tasks such as recall for recollection. As this dissertation demonstrated through the high relation of familiarity with recall, this method is not a process-pure estimate and different conclusions can be made when parameters are directly estimated. Therefore, future research using similar procedures in a paradigm where recollection is higher on average, will prove useful in confirming the conclusions that recollection is an important predictor of cognitive abilities. Of course, the variability of lures could also play a larger role in conditions or paradigms where recollection estimates are higher.

Another possibility is that younger adults do not vary in their recollective abilities as much as older adults. There has been considerable support for the idea that older adults as a group have a deficit in recollection (Jennings & Jacoby, 1997; Koen & Yonelinas, 2014). If some older adults are experiencing larger declines in recollection, there may be greater variability of recollection in older adults, allowing for the possibility that recollection is more useful as a measure of cognitive ability. On the other hand, a deficit in recollection arguably forces older adults to rely on familiarity more, therefore potentially making it even more predictive of abilities such as gF and recall.

The CDP was proposed as a combination of the UVSD and DPSD in part because of the accumulation of evidence that dual-process models have proven useful (Wixted & Mickes, 2012). In proposing the CDP, Wixted and Mickes (2012) wrote, “despite its utility in some domains, the concept of memory strength seems like a woefully inadequate construct to capture the richness of memory.” (*p*. 1025). For example, the remember-know task has become popular in neurocognitive research and has purportedly identified different brain regions responsible for the two types of judgments (e.g., Kim, 2010). The UVSD does not offer estimates of separate parameters from the R/K task and thus the difference in remember and know judgments are a function of strength, which does not clearly explain why different brain regions would be activated. As shown here, DPSD parameters are reliable, exhibit convergent validity, and are relatively highly correlated with each other when derived from different estimation methods. The DPSD is not a perfect model; as evidenced by the higher-than-expected relation of familiarity to recall, it likely oversimplifies cognitive processes that are interrelated. However, DPSD parameter estimates appear to be the only recognition memory model that isolate two specific processes that are both reliable and predict other cognitive abilities.

5.3 Implications for Individual Differences Research in Recognition Memory

In addition to providing a description and comparison of recognition memory models, this dissertation aimed to advance research on individual differences in LTM and specifically recognition memory. For the purposes of individual differences research, whether any of the models are more useful for predicting cognitive abilities, then simply using recognition performance statistic that is not directly tied to a recognition memory model (although every performance metric makes implicit assumptions about what recognition memory is testing and the shape of the ROC). Overall memory performance, as measured by Corrected HR, significantly predicted gF, with a standardized path coefficient of .28, and was unrelated to PS. None of the recognition memory models performed better than Corrected HR, based on overlapping confidence interval of variance explained in gF.

In this sample, *d′* was roughly equivalent to Corrected HR, ($r = .93$) and the V_{old} parameter was not predictive of either gF nor PS. Therefore, there was no benefit to estimating UVSD parameters from the perspective of providing further insight into individual differences, as a much simpler calculation of Corrected HR captured approximately all the same variance. This does not mean the model is incorrect because there is nothing inconsistent with a single memory strength model's key parameter being highly correlated with Corrected HR. On the other hand, there is a potential benefit of utilizing DPSD parameters in individual differences research, as these parameters were highly reliable and only moderately correlated, allowing for the possibility that different abilities may be captured by recollection and familiarity. In practice, the DPSD models estimated from both the ROC and R/K judgments explained more variance from gF than the Corrected HR, but there was considerable overlap of the confidence intervals. Further studies of individual differences that have larger samples or test different abilities have potential to find differences in relationships between recollection and familiarity and Corrected HR.

Another finding with implications for individual differences research is that domain-general PS was not greatly related to individual differences in LTM in younger adults. Neither recall nor any parameter from any recognition memory parameter was related to PS. This may have been partially due to PS being better characterized by multiple domains (e.g., Hale & Myerson, 1996; Lawrence et al., 1998). The lone verbal PS task, the animal judgment task, exhibited higher correlations with many of the memory parameters (which were based on a verbal memory test) than the other PS tasks (Table 4.2). For example, the animal PS tasks correlation with familiarity was $r = -.18$ and for d' was $r = -.17$, whereas both those parameters did not significantly correlate with the dot and shape PS tasks. However, these correlation coefficients with the animal judgment task were still low in magnitude. PS may be a more useful variable in cognitive aging research than in research that only assesses individual differences in younger adults (Salthouse, 1996)

In addition, verbal fluency, which was posited to be related to familiarity precisely because of its verbal nature, was not related to any of the recognition memory parameters, nor was it related to any of the other cognitive abilities. This construct, at least as assessed here, does not appear to be useful towards understanding individual differences in younger adults. This was a surprising finding, because several studies had found relatively large contributions of a verbal fluency factor with both recognition memory (Hedden et al., 2005) and general LTM, as well as with both gF and PS (Unsworth, 2019). Perhaps, the 60 seconds given here was not enough to find adequate differences, as Hedden et al. (2005) gave participants 90 seconds to list words, whereas the other details of the method were similar. Another possibility is that the unsupervised nature of the online session did not allow participants to fully understand or put effort into the task. As with PS, this variable may also become more relevant as an individual difference in an older or more diverse sample (i.e., not all undergraduate students at a selective university).

5.4 Limitations and future directions

This dissertation offered a novel analysis of recognition memory models from the perspective of individual differences. Although several conclusions can be made from the study, many of the analyses failed to find conclusive differences between models. A combination of factors contributed to the equivocal findings. SEM can often capitalize on chance, meaning that large samples are required, and conclusions may be considered tentative without replication. For this dissertation, this issue was exacerbated by the high correlations between the parameters from different recognition memory models (Table 4.3). The high correlations between parameters demonstrates that despite the differences between the models, they are calculated from the same data and share a basis in signal detection theory. For the purposes of model comparison, the correlated parameters make finding differences between the models difficult because much of the shared variance between a parameter and cognitive abilities was likely also shared between other models' parameters.

Another factor potentially impacting the findings was that the Remember-rate was relatively low, as was recollection calculated from ROCs. The remember rate of .19 for old items was lower than many other studies of young adults. In a representative study, Prull et al. (2006) found a Remember-rate of .55 for young adults with 60 studied words in a block and 30 lures. However, other studies have shown low Remember-rates with long lists of words and no deep encoding instructions. Yonelinas (2001) also found a Remember-rate of .19 for a shallow encoding condition of 80 words (with 80 deeply encoded words). There are numerous factors that have been demonstrated to effect recollection, most notably list length, method of encoding, and characteristics of the words such as frequency and relatedness (Yonelinas, 2002). The lists used in both studies were not prone to a large amount of recollection primarily due to length and due to shallow encoding. The words also were relatively common unrelated words, which are not the ideal conditions for strong performance in recognition tasks, nor high remember tasks (e.g., Yonelinas, 2002). Finally, the instructions chosen for the R/K task were based on the conservative instructions from Rotello et al. (2005). These conservative instructions are best for emphasizing that remember decisions are based on recollections because they specify that a specific detail about the context must be remembered, but they could lead to an under-estimate of recollection if participants take an overly conservative approach.

Another factor that may have influenced lower levels of recollection was potential confusion about the combination of confidence and remember-know decisions. According to the DPSD, the vast majority of successful recollections should be associated with both a '6' confidence and a 'remember' response. Rather than responding with high confidence and Remember participant may have chosen one but not the other as they want to perform the task correctly, and researchers may not want them to respond the same way every time. By not choosing either '6' confidence or 'r' responses for some recollections the rate of both these responses would be depressed. Across participants, 29% of old items received the highest confidence rating, and 22% were "remembered." 83.33% of remember responses to old items were associated with the highest confidence. This was nearly identical to the proportion of remember responses that were associated with highest confidence, 82.77%, in the conservative condition of the Rotello et al. (2005) study despite that study using a different confidence scale, where confidence levels 2-6 were considered 'old' judgments. These findings are consistent with other studies demonstrating that, when asked, participants will not always respond with the highest confidence for remember judgments, although they do for most of them (Ingram et al., 2012). It is unclear whether the lack of unity of high confidence and remember is due to task demands or a flaw with the DPSD. An interesting comparison for future study is whether participants show different patterns of responding when both R/K and confidence decisions are asked for versus when either judgment is asked for alone.

Another potential flaw with the R/K task with confidence ratings was the incentivization of responding ‘new’ (1-3 confidence), since this response advanced to the next memory trial, rather than asking for a R/K judgment. Although the time required for the R/K response was brief, over the course of 240 test trials participants may have been biased to skip that trial when possible. Table 4.1 shows that 19% of old items received a response of ‘3’ compared to 15% that received a ‘4’. In comparison, 29% of new items received a ‘3’ compared to 12% that received a ‘4’. These data show potential evidence that the incentive to not have to make another choice (and delay progression on the test) may have led to a slight bias to respond ‘3’ in cases when the participant was uncertain, perhaps even if the participants had some memory for the item. Future implementations of this paradigm should consider adding an additional response screen for ‘new’ responses as well to not incentivize skipping the R/K judgment.

Given the results suggesting that the CDP was not a reliable model for explaining individual differences in cognitive abilities, the R/K decision and confidence combination did not prove to be necessary. The UVSD only factors in confidence ratings, and the DPSD either R/K or confidence, but not both. Given the closeness in performance of these two models in Study 2 and the potential flaws in the R/K with confidence paradigm that have been outlined, a similar study that only asked for confidence ratings might prove useful in further comparing those two models. Such a study would avoid any potential noise that is introduced by the R/K judgments and decrease the time needed for the task.

It should be noted that only one version of the CDP was tested. The CDP does not necessitate that there is unequal variance between lures and targets for recollection and equal variance for familiarity. Alternate versions of the CDP could be calculated with unequal variance of familiarity and/or equal variance of recollection. However, this study calculated the same version of the CDP that was calculated by Wixted and Mickes' (2010) analysis of data from Rotello et al. (2005), which was, to my knowledge, the only study to calculate parameters for the CDP. CDP theorists would have to clarify why a different version would be appropriate for these data. Of course, there are also numerous other models of recognition memory and LTM, that could prove to be useful in assessing individual differences, but I chose to focus on two models that have often been pitted against each other, the UVSD and the DPSD, as well as the CDP which has been proposed to improve on both of those models.

Finally, all conclusions from this study are limited to the verbal domain. There is some evidence that verbal and visuospatial tasks assess different LTM abilities (Siedlecki, 2005). As mentioned, the same is true of other cognitive constructs, such as PS (Hale & Myerson, 1996), so verbal-specific versions of constructs may be more related to the verbal abilities assessed here. The conclusions about reliability and validity of the various models may not apply to stimuli other than words, particularly if those stimuli are visuospatial. Future studies ought to clarify if the DPSD and UVSD models exhibit reliability and convergent validity in the visuospatial domain.

5.5 Conclusion

As with other methods of model comparison and validation, the evidence from a single study of individual differences cannot on its own establish a model's validity nor establish the optimal model. However, as this dissertation showed, individual difference research in conjunction with experimental research can make major contributions towards understanding recognition and comparing recognition memory models. The results of this studied established that the CDP was unreliable, a major shortcoming of the model. Overall, the data provided support for both the UVSD and the DPSD and largely supported the hypotheses of the DPSD, including that estimates of recollection and familiarity both predicted individual differences in other cognitive abilities, including gF. These findings demonstrate the potential utility of the DPSD model for individual differences. Further studies of individual differences would be useful in contributing towards understanding which recognition memory models best capture the memory abilities that underlie performance in recognition memory.

References

Aichele, S., Rabbitt, P., & Ghisletta, P. (2015). Life span decrements in fluid intelligence and processing speed predict mortality risk. *Psychology and Aging*, *30*(3), 598–612. https://doi.org/10.1037/pag0000035

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological bulletin*, *103*(3), 411.

Babyak, M. A. (2004). What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. *Psychosomatic Medicine*, *66*(3), 411–421.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior research methods*, *39*(3), 445-459.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, *107*(2), 238.

Berry, C. J., Shanks, D. R., & Henson, R. N. A. (2008). A single-system account of the relationship between priming, recognition, and fluency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 97–111. https://doi.org/10.1037/0278-7393.34.1.97

Bird, C. M., Papadopoulou, K., Ricciardelli, P., Rossor, M. N., & Cipolotti, L. (2003). Test-retest reliability, practice effects and reliable change indices for the recognition memory test. *British Journal of Clinical Psychology*, *42*(4), 407–425. https://doi.org/10.1348/014466503322528946

Blankenship, T. L., O'Neill, M., Ross, A., & Bell, M. A. (2015). Working memory and recollection contribute to academic achievement. *Learning and Individual Differences*, *43*, 164–169. https://doi.org/10.1016/j.lindif.2015.08.020

Bollen, K. A. (1980). Issues in the Comparative Measurement of Political Democracy. *American Sociological Review*, *45*(3), 370–390. https://doi.org/10.2307/2095172

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.2307/2286841

Brown, S., & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 11–21. https://doi.org/10.3758/BF03195493

Calvin, C. M., Deary, I. J., Fenton, C., Roberts, B. A., Der, G., Leckenby, N., & Batty, G. D. (2011). Intelligence in youth and all-cause-mortality: Systematic review with meta-analysis. *International Journal of Epidemiology*, *40*(3), 626–644.

Cattell, R. B. (1973). *Culture-fair intelligence test.* Institute for personality and ability testing.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American journal of mental deficiency*, *86*(2), 127–137.

Cohen, A. L., Rotello, C. M., & Macmillan, N. A. (2008). Evaluating models of remember-know judgments: Complexity, mimicry, and discriminability. *Psychonomic Bulletin & Review*, *15*(5), 906–926. https://doi.org/10.3758/PBR.15.5.906

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*(2), 163–184. https://doi.org/10.1016/S0160-2896(01)00096-4

Craik, F. I. M., & McDowd, J. M. (1997). *Age Differences in Recall and Recognition*. 6.

Daselaar, S. M., Fleck, M. S., & Cabeza, R. (2006). Triple Dissociation in the Medial Temporal Lobes: Recollection, Familiarity, and Novelty. *Journal of Neurophysiology*, *96*(4), 1902–1911. https://doi.org/10.1152/jn.01029.2005

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*(1), 13–21.

Dede, A. J. O., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2013). Hippocampal damage impairs recognition memory broadly, affecting both parameters in two prominent models of memory. *Proceedings of the National Academy of Sciences*, *110*(16), 6577–6582. https://doi.org/10.1073/pnas.1304739110

Dobbins, I. G., Khoe, W., Yonelinas, A. P., & Kroll, N. E. A. (2000). Predicting individual false alarm rates and signal detection theory: A role for remembering. *Memory & Cognition*, *28*(8), 1347–1356. https://doi.org/10.3758/BF03211835

Dunn, J. C. (2008). The dimensionality of the remember-know task: A state-trace analysis. *Psychological Review*, *115*(2), 426–446. https://doi.org/10.1037/0033-295X.115.2.426

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The Medial Temporal Lobe and Recognition Memory. *Annual Review of Neuroscience*, *30*(1), 123–152. https://doi.org/10.1146/annurev.neuro.30.051606.094328

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of experimental psychology: General*, *128*(3), 309.

Ghisletta, P., Rabbitt, P., Lunn, M., & Lindenberger, U. (2012). Two thirds of the age-based changes in fluid and crystallized intelligence, perceptual speed, and memory in adulthood are shared. *Intelligence*, *40*(3), 260–268. https://doi.org/10.1016/j.intell.2012.02.008

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 500–513. https://doi.org/10.1037/0278-7393.25.2.500

Gonsalves, B. D., Kahn, I., Curran, T., Norman, K. A., & Wagner, A. D. (2005). Memory Strength and Repetition Suppression: Multimodal Imaging of Medial Temporal Cortical Contributions to Recognition. *Neuron*, *47*(5), 751–761. https://doi.org/10.1016/j.neuron.2005.07.013

Goodhew, S. C., & Edwards, M. (2019). Translating experimental paradigms into individual-differences research: Contributions, challenges, and practical recommendations. *Consciousness and Cognition*, *69*, 14–25. https://doi.org/10.1016/j.concog.2019.01.008

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hale, S., & Myerson, J. (1996). *Experimental evidence for differential slowing in the lexical and nonlexical domains. Aging, Neuropsychology, and Cognition, 3*(2), 154-165.

Healy, M. R., Light, L. L., & Chung, C. (2005). Dual-Process Models of Associative Recognition in Young and Older Adults: Evidence From Receiver Operating Characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 768–788. https://doi.org/10.1037/0278-7393.31.4.768

Heathcote, A. (2003). Item Recognition Memory and the Receiver Operating Characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1210–1230. https://doi.org/10.1037/0278-7393.29.6.1210

Heathcote, A., Raymond, F., & Dunn, J. (2006). Recollection and familiarity in recognition memory: Evidence from ROC curves. *Journal of Memory and Language*, *55*(4), 495–

514. https://doi.org/10.1016/j.jml.2006.07.001

Hedden, T., Lautenschlager, G., & Park, D. C. (2005). Contributions of Processing Ability and Knowledge to Verbal Memory Tasks across the Adult Life-Span. *The Quarterly Journal of Experimental Psychology Section A*, *58*(1), 169–190. https://doi.org/10.1080/02724980443000179

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hintzman, D. L. (2011). Research Strategy in the Study of Memory: Fads, Fallacies, and the Search for the “Coordinates of Truth.” *Perspectives on Psychological Science*, *6*(3), 253–271. https://doi.org/10.1177/1745691611406924

Hintzman, D. L., & Caulton, D. A. (1997). Recognition Memory and Modality Judgments: A Comparison of Retrieval Dynamics. *Journal of Memory and Language*, *37*(1), 1–23. https://doi.org/10.1006/jmla.1997.2511

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of educational psychology*, *57*(5), 253.

Howard, M. W., Bessette-Symons, B., Zhang, Y., & Hoyer, W. J. (2006). Aging selectively impairs recollection in recognition memory for pictures: Evidence from modeling and receiver operating characteristic curves. *Psychology and Aging*, *21*(1), 96–106. https://doi.org/10.1037/0882-7974.21.1.96

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Ingram, K. M., Mickes, L., & Wixted, J. T. (2012). Recollection can be weak and familiarity can be strong. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 325–339. https://doi.org/10.1037/a0025483

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology. General*, *110*(3), 306–340. https://doi.org/10.1037/0096-3445.110.3.306

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General*, *138*(2), 291–306. https://doi.org/10.1037/a0015525

Jennings, J. M., & Jacoby, L. L. (1997). An opposition procedure for detecting age-related deficits in recollection: telling effects of repetition. *Psychology and aging*, *12*(2), 352.

Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences*. Elsevier.

Kim, H. (2010). Dissociating the roles of the default-mode, dorsal, and ventral networks in episodic memory retrieval. *NeuroImage*, *50*(4), 1648–1657. https://doi.org/10.1016/j.neuroimage.2010.01.051

King, D. R., de Chastelaine, M., & Rugg, M. D. (2018). Recollection-related increases in functional connectivity across the healthy adult lifespan. *Neurobiology of Aging*, *62*, 1–19. https://doi.org/10.1016/j.neurobiolaging.2017.09.026

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. Guilford publications.

Koen, J. D., & Yonelinas, A. P. (2014). The Effects of Healthy Aging, Amnestic Mild Cognitive Impairment, and Alzheimer's Disease on Recollection and Familiarity: A Meta-Analytic Review. *Neuropsychology Review*, *24*(3), 332–354. https://doi.org/10.1007/s11065-014-9266-5

Koen, J. D., & Yonelinas, A. P. (2016). Recollection, not familiarity, decreases in healthy ageing: Converging evidence from four estimation methods. *Memory*, *24*(1), 75–88. https://doi.org/10.1080/09658211.2014.985590

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Lawrence, B., Myerson, J., & Hale, S. (1998). Differential Decline of Verbal and Visuospatial Processing Speed Across the Adult Life Span. *Aging, Neuropsychology, and Cognition*, *5*(2), 129–146. https://doi.org/10.1076/anec.5.2.129.600

Lombardi, M. G., Perri, R., Fadda, L., Caltagirone, C., & Carlesimo, G. A. (2018). Forgetting of the recollection and familiarity components of recognition in patients with amnestic mild cognitive impairment. *Journal of Neuropsychology*, *12*(2), 231–247. https://doi.org/10.1111/jnp.12114

Long, D. L., Prat, C., Johns, C., Morris, P., & Jonathan, E. (2008). The importance of knowledge in vivid text memory: An individual-differences investigation of recollection and familiarity. *Psychonomic Bulletin & Review*, *15*(3), 604–609. https://doi.org/10.3758/PBR.15.3.604

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

https://doi.org/10.3758/BF03204766

Malmi, R. A., Underwood, B. J., & Carroll, J. B. (1979). The interrelationships among some associative learning tasks. *Bulletin of the Psychonomic Society*, *13*(3), 121–123. https://doi.org/10.3758/BF03335032

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological review*, *87*(3), 252-271.

McCabe, D. P., Roediger, H. L., & Karpicke, J. D. (2011). Automatic processing influences free recall: Converging evidence from the process dissociation procedure and remember-know judgments. *Memory & Cognition*, *39*(3), 389–402. https://doi.org/10.3758/s13421-010-0040-5

McCabe, D. P., Roediger III, H. L., McDaniel, M. A., & Balota, D. A. (2009). Aging reduces veridical remembering but increases false remembering: Neuropsychological test correlates of remember–know judgments. *Neuropsychologia*, *47*(11), 2164-2173.

McCabe, D. P., Roediger, H. L., McDaniel, M. A., Balota, D. A., & Hambrick, D. Z. (2010). The relationship between working memory capacity and executive functioning: Evidence for a common executive attention construct. *Neuropsychology*, *24*(2), 222–243. https://doi.org/10.1037/a0017619

Mickes, L., Seale-Carlisle, T. M., & Wixted, J. T. (2013). Rethinking familiarity: Remember/Know judgments in free recall. *Journal of Memory and Language*, *68*(4), 333–349. https://doi.org/10.1016/j.jml.2013.01.001

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

Prull, M. W., Dawes, L. L. C., Martin, A. M., Rosenberg, H. F., & Light, L. L. (2006). Recollection and familiarity in recognition memory: Adult age differences and neuropsychological test correlates. *Psychology and Aging*, *21*(1), 107–118. https://doi.org/10.1037/0882-7974.21.1.107

Quamme, J. R., Yonelinas, A. P., Widaman, K. F., Kroll, N. E. A., & Sauvé, M. J. (2004). Recall and recognition in mild hypoxia: Using covariance structural modeling to test competing theories of explicit memory. *Neuropsychologia*, *42*(5), 672–691. https://doi.org/10.1016/j.neuropsychologia.2003.09.008

Raven, J. (2000). The Raven's progressive matrices: change and stability over culture and time. *Cognitive psychology*, *41*(1), 1-48.

Revelle W (2020). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of statistical software*, *48*(2), 1-36.

Rotello, C. M., & Macmillan, N. A. (2006). Remember–know models as decision strategies in two experimental paradigms. *Journal of Memory and Language*, *55*(4), 479–494. https://doi.org/10.1016/j.jml.2006.08.002

Rhodes, S., Greene, N. R., & Naveh-Benjamin, M. (2019). Age-related differences in recall and recognition: A meta-analysis. *Psychonomic Bulletin & Review*, *26*(5), 1529–1547. https://doi.org/10.3758/s13423-019-01649-y

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367. https://doi.org/10.1037/0033-295X.107.2.358

Rotello, C. M., Macmillan, N. A., Reeder, J. A., & Wong, M. (2005). The remember response:

Subject to bias, graded, and not a process-pure indicator of recollection. *Psychonomic Bulletin & Review*, *12*(5), 865–873. https://doi.org/10.3758/BF03196778

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452–467. https://doi.org/10.3758/s13423-018-1558-y

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological review*, *103*(3), 403.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262–274.

Sheppard, L. D., & Vernon, P. A. (2008). Intelligence and speed of information-processing: A review of 50 years of research. *Personality and Individual Differences*, *44*(3), 535–551. https://doi.org/10.1016/j.paid.2007.09.015

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

Siedlecki, K. L. (2007). Investigating the structure and age invariance of episodic memory across the adult lifespan. *Psychology and Aging*, *22*(2), 251–268. https://doi.org/10.1037/0882-7974.22.2.251

Slotnick, S. D., & Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory & Cognition*, *33*(1), 151–170. https://doi.org/10.3758/BF03195305

Smith, D. G., & Duncan, M. J. J. (2004). Testing Theories of Recognition Memory by Predicting Performance Across Paradigms. *Journal of Experimental Psychology: Learning,* *Memory, and Cognition*, *30*(3), 615–625. https://doi.org/10.1037/0278-7393.30.3.615

Spanton, R. W., & Berry, C. J. (2020). The unequal variance signal-detection model of recognition memory: Investigating the encoding variability hypothesis. *Quarterly Journal of Experimental Psychology*, *73*(8), 1242–1260. https://doi.org/10.1177/1747021820906117

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate behavioral research*, *25*(2), 173-180.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric monographs*.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne*, *26*(1), 1–12. https://doi.org/10.1037/h0080017

Uncapher, M. R., & Rugg, M. D. (2005). Encoding and the Durability of Episodic Memory: A Functional Magnetic Resonance Imaging Study. *Journal of Neuroscience*, *25*(31), 7260–7267. https://doi.org/10.1523/JNEUROSCI.1641-05.2005

Underwood, B. J. (1975). Individual differences as a crucible in theory construction. *American Psychologist*, *30*(2), 128–134. https://doi.org/10.1037/h0076759

Unsworth, N. (2010). On the division of working memory and long-term memory and their relation to intelligence: A latent variable approach. *Acta Psychologica*, *134*(1), 16–28. https://doi.org/10.1016/j.actpsy.2009.11.010

Unsworth, N. (2019). Individual differences in long-term memory. *Psychological Bulletin*, *145*(1), 79–139. https://doi.org/10.1037/bul0000176

Unsworth, N., & Brewer, G. A. (2009). Examining the relationships among item recognition, source recognition, and recall from an individual differences perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1578–1585.

https://doi.org/10.1037/a0017255

Wang, W., & Yonelinas, A. P. (2012). Familiarity is related to conceptual implicit memory: An examination of individual differences. *Psychonomic Bulletin & Review*, *19*(6), 1154–1164. https://doi.org/10.3758/s13423-012-0298-7

Westerman, D. L., Miller, J. K., & Lloyd, M. E. (2003). Change in perceptual form attenuates the use of the fluency heuristic in recognition. *Memory & Cognition*, *31*(4), 619–629. https://doi.org/10.3758/BF03196102

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152–176. https://doi.org/10.1037/0033-295X.114.1.152

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review*, *117*(4), 1025–1054. https://doi.org/10.1037/a0020874

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*, *11*(4), 616–641. https://doi.org/10.3758/BF03196616

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational and Psychological Measurement*, *73*(6), 913–934. https://doi.org/10.1177/0013164413495237

Wrulich, M., Brunner, M., Stadler, G., Schalke, D., Keller, U., & Martin, R. (2014). Forty years on: Childhood intelligence predicts health in middle adulthood. *Health Psychology*, *33*(3), 292–296.

Yonelinas, A. P. (1994). Receiver-Operating Characteristics in Recognition Memory: Evidence for a Dual-Process Model. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, *20*(6), 1341–1354.

Yonelinas, A. P. (2001). Components of episodic memory: The contribution of recollection and familiarity. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *356*(1413), 1363–1374. https://doi.org/10.1098/rstb.2001.0939

Yonelinas, A. P., & Jacoby, L. L. (1995). The relation between remembering and knowing as bases for recognition: Effects of size congruency. *Journal of memory and language*, *34*(5), 622-643

Yonelinas, A. P., & Jacoby, L. L. (2012). The process-dissociation approach two decades later: Convergence, boundary conditions, and new directions. *Memory & Cognition*, *40*(5), 663–680. https://doi.org/10.3758/s13421-012-0205-5

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*(5), 800–832. https://doi.org/10.1037/0033-2909.133.5.800

Zerr, C. L., Berg, J. J., Nelson, S. M., Fishell, A. K., Savalia, N. K., & McDermott, K. B. (2018). Learning Efficiency: Identifying Individual Differences in Learning Rate and Retention in Healthy Adults. *Psychological Science*, *29*(9), 1436–1450. https://doi.org/10.1177/0956797618772540

<u>Appendix A</u>

Additional Reliability Statistics for Recognition Memory data

Table A.1 Reliability for Model Parameters with extreme outliers and non-Native English speakers' data removed

Parameter	*n*	Pearson *r*	ICC
DPSD model R/K	60		
Recollection A		.87	.85
Recollection B		(.78, .93)	(.78, .90)
Familiarity A		.81	.79
Familiarity B		(.68, .89)	(.69, .85)
CDP model	54		
Mean-of-R A		.54	.58
Mean-of-R A		(.30, .71)	(.40, .71)
Mean-of-F A		.23	.17
Mean-of-F B		(-0.05, .50)	(-.05, .37)
SD-of-R A		.59	.60
SD-of-R B		(.39, .74)	(.43, .72)
UVSD model	53		
d′ A		.81	.78
d′ B		(.64, .91)	(.67, .85)
V_{old} A		.46	.45
V_{old} B		(.26, .63)	(.25, .62)
DPSD model ROC	60		
Recollection A		.78	.77
Recollection B		(.62, .88)	(.67, .85)
Familiarity A		.75	.70
Familiarity B		(.58, .88)	(.58, .80)

Note. 95% confidence intervals in parentheses.

Table A.2 Reliability for CDP Model Criteria

Parameter	Pearson *r*	ICC
Confidence 6	.48 (.25, .65)	.41 (.21, .57)
Confidence 5	.57 (.37 .72)	.54 (.37, .68)
Confidence 4	.68 (.51, .79)	.49 (.31, .64)
Confidence 3	.64 (.47, .77)	.70 (.57, .79)
Confidence 2	.71 (.55, .82)	.60 (.55, .73)
Remember Criterion	.63 (.44, .76)	.70 (.57, .79)

Note. 95% confidence intervals in parentheses. Based on data with extreme outliers removed (n = 59)

Table A.3 Reliability for UVSD Model Criteria

Parameter	Pearson *r*	ICC
Confidence 6	.60 (.46, .74)	.58 (.41, .72)
Confidence 5	.70 (.52, .87)	.64 (.48, .76)
Confidence 4	.74 (.63, .89)	.66 (.50, .78)
Confidence 3	.74 (.59, .85)	.68 (.54, .79)
Confidence 2	.65 (.50, .78)	.55 (.36, .69)

Note. 95% confidence intervals in parentheses. Based on data with extreme outliers removed (n = 50).

Table A.4 Reliability for DPSD ROC Model Criteria

Parameter	Pearson r	ICC
Confidence 6	.76 (.54, .89)	.75 (.64, .84)
Confidence 5	.73 (.60, .86)	.73 (.60, .81)
Confidence 4	.78 (.69, .86)	.67 (.53, .77)
Confidence 3	.73 (.61, .84)	.58 (.41, .70)
Confidence 2	.56 (.37, .72)	.55 (.39, 69)

Note. 95% confidence intervals in parentheses. Based on data with extreme outliers removed (n = 59).

Appendix B

Verbal Fluency Correlations

Table B.1. *Verbal fluency Correlations with Model parameters and cognitive measures*

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. Recollect ROC																					
2. Famil ROC	.29																				
3. Recollect R/K	.62	.42																			
4. Famil R/K	.52	.75	.15																		
5. mean-of-R	.48	.51	.55	.42																	
6. mean-of-F	.22	.13	.06	.20	-.59																
7. SD-of-R	.52	.36	.65	.19	.36	.25															
8. d′UVSD	.80	.74	.65	.76	.63	.22	.60														
9. V_{old} UVSD	.65	.18	.42	.33	.32	.22	.69	.66													
10. RT dot	.03	.05	.07	-.01	.02	.01	-.00	.03	-.05												
11. RT animal	-.10	-.15	-.07	-.18	-.06	-.08	-.10	-.17	-.09	.42											
12. RT shape	-.02	.01	-.02	-.04	.04	-.09	-.05	-.06	-.10	.57	.47										
13. RAPM	.11	.36	.13	.31	.27	-.02	.22	.30	.07	-.21	-.09	-.16									
14. Number Ser	.13	.25	.16	.18	.17	.06	.24	.23	.04	-.14	-.16	-.15	.53								
15. Cattell	.14	.28	.22	.24	.21	.05	.21	.27	.06	-.09	-.08	-.04	.46	.39							
16. 25word Recall	.25	.27	.25	.24	.28	-.01	.25	.33	.23	-.12	-.18	-.17	.30	.23	.17						
17. R/K Recall 1	.44	.44	.39	.41	.41	.04	.32	.52	.27	-.05	-.18	-.16	.25	.21	.30	.42					
18. R/K Recall 2	.53	.33	.40	.37	.34	.14	.34	.55	.36	.04	-.12	-.12	.11	.14	.14	.29	.44				
19. Fruit words	.17	.11	**.18**	.09	.17	-.04	.13	.17	.12	-.01	**-.25**	-.10	.05	**.20**	.02	**.24**	**.20**	**.24**			
20. Furniture word	.05	.11	.05	.07	-.01	.12	.05	.10	-.00	.00	-.13	.01	.06	**.22**	.03	.14	.04	.14	**.47**		
21. M words	.06	-.02	-.02	-.08	-.04	.01	-.00	-.05	-.03	-.09	**-.26**	-.00	-.01	.03	.04	.09	.01	.08	**.38**	**.52**	
22. S words	.02	.00	.03	-.02	.06	-.04	.01	.03	.04	-.11	**-.28**	-.10	-.02	.12	-.01	.13	.02	.05	**.44**	**.36**	**.58**

Note. Significant (p < .05) correlations with verbal fluency measures are in bold. All absolute values greater than or equal to .15 were significant, p < .05. All absolute values greater than or equal to .19 were significant p < .01. Recollection, Famil = Familiarity, Cattell = Cattell Culture Fair Matrices, RAPM = Raven's Advanced Progressive Matrices.

Appendix C

Multiple Linear Regression Results

Table C.1. *Regression results using gF as the criterion*

Variable	*B*	*SE*	ß	ß 95% CI	*p*	*Model R^2*
			DPSD R/K model			
						.162
Intercept	-0.90	0.16			< .001	
Recollect	1.19	0.43	0.19	[0.06, 0.32]	.006	
Familiar	2.30	0.47	0.33	[0.20, 0.46]	< .001	
			CDP model			
						.176
Intercept	-0.63	0.17			< .001	
Mean Recollect	0.34	0.08	0.46	[0.25, 0.68]	< .001	
Mean Familiar	0.27	0.10	0.29	[0.08, 0.50]	.007	
SD-of-R	0.09	0.12	0.07	[-0.11, 0.25]	.447	
			UVSD model			
						.184
Intercept	-0.99	0.17			< .001	
discrimin	0.76	0.12	0.55	[0.38, 0.72]	<.001	
Vold	-0.71	0.26	-0.25	[-0.42, -0.07]	.006	
			DPSD ROC model			
						.179
Intercept	-0.76	0.15			< .001	
Recollect	0.62	0.46	0.09	[-0.04, 0.23]	.180	
Familiar	0.93	0.17	0.39	[0.25, 0.52]	< .001	

Note. gF scores were participants' standardized predicted scores on the gF factor. The SD terms (SD-of-R and Vold) were centered, all other scores are in their original metric. ß indicates the standardized regression weights. CI = Confidence Interval

Table C.2. *Regression results using Recall as the criterion*

Variable	*B*	*SE*	ß	ß 95% CI	*p*	*Model* R^2
			DPSD R/K model			
						.360
Intercept	-1.30	0.14			< .001	
Recollect	2.41	0.37	0.38	[0.26, 0.50]	< .001	
Familiar	2.87	0.42	0.41	[0.29, 0.53]	< .001	
			CDP model			
						.387
Intercept	-1.09	0.15			< .001	
Mean Recollect	0.58	0.07	0.78	[0.60, 0.97]	< .001	
Mean Familiar	0.50	0.08	0.54	[0.36, 0.72]	< .001	
SD-of-R	-0.02	0.10	-0.02	[-0.17, 0.14]	.813	
			UVSD model			
						.378
Intercept	-1.24	0.15			< .001	
discrimin	0.94	0.11	0.68	[0.53, 0.84]	< .001	
Vold	-0.34	0.22	-0.11	[-0.27, 0.04]	.135	
			DPSD ROC model			
						.378
Intercept	-1.22	0.13			< .001	
Recollect	2.58	0.40	0.39	[0.27, 0.51]	< .001	
Familiar	0.91	0.15	0.38	[0.26, 0.49]	< .001	

Note. Recall scores were participants' standardized predicted scores on the Recall factor. The SD terms (SD-of-R and Vold) were centered, all other scores are in their original metric. ß indicates the standardized regression weights. CI = Confidence Interval