



different measurement models of memory and find that the type of relation between memory and judgment changes as the function of the type of memory model that one adopts. Finally, combined data across both experiments reveal that how the to-be-judged items are distributed plays a role in judgments and that retrieval ability, specifically, semantic memory, is predictive of probability judgments. Taken together, we argue that the ability to retrieve from LTM plays a critical role in judging the likelihood of an event occurring.

GROUNDING JUDGMENT PHENOMENA IN MEMORY: EXAMINING THE  
ROLE OF RETRIEVAL IN THE ESTIMATION OF EVENTS

by

Rosalind Nguyen

Thesis submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park, in partial fulfillment  
of the requirements for the degree of  
Masters of Science  
2018

Advisory Committee:  
Professor Michael Dougherty, Chair  
Associate Professor Robert Slevc  
Assistant Professor Edward Bernat

© Copyright by  
Rosalind Nguyen  
2018

## Dedication

I would like to dedicate this to my family and friends, who have provided their love and support to me over the years. To Kelly Segura, my dearest friend, who took time out of her life to provide me with the care and emotional support I needed from afar. And finally, to my parents, Tuyen and Lien Nguyen, who immigrated to this country. Their sacrifices have afforded me with the opportunities I have today.

## Acknowledgements

I would like to give thanks and acknowledgments to the many undergraduate research assistants that assisted in collecting data for the studies outlined in this project. I would also like to thank my lab mates, Alison Robey and David Ampofo, for their advice and guidance throughout the year. Finally, I would like to thank Dr. Michael Dougherty for welcoming me into his lab and advising me on my work throughout the years.

# Table of Contents

Dedication .....	ii
Acknowledgements .....	iii
Table of Contents .....	iv
List of Tables .....	vii
List of Figures .....	viii
Chapter 1: Grounding Judgment Phenomena in Memory .....	1
Probability Judgments and Subadditivity .....	1
Judgments and Hypothesis Generation .....	3
The Role of Memory in Hypothesis Generation .....	5
Chapter 2: Current Study .....	7
Questions and Hypotheses .....	9
<i>Question 1: Is there evidence of the comparison process for probability judgments and lack of one for frequency judgments?</i> .....	9
<i>Question 2: What is the relation between each judgment type and memory ability?</i> .....	9
Experiments .....	10
Experiment 1 .....	10
Experiment 2 .....	11
Combined Data Analysis .....	11
Chapter 3: General Data Analysis Methods and Approach .....	11
Model Comparison .....	12
Latent Variables Approach .....	13
Chapter 4: Experiment 1 .....	13
Method .....	14
Participants .....	14
Design .....	14
Procedure .....	15
Measures .....	15
Assessing Probability and Frequency Judgments .....	16
Assessing Working Memory .....	18
Assessing Retrieval Ability .....	21
Assessing Recall of the To-Be-Judged Items .....	23
Analysis .....	23
Coding of Variables & Constructs .....	23
Missing Data .....	24
Outliers and Robustness Analysis .....	24
Software .....	25
Results .....	25
Final Analysis Sample and Demographics .....	25
Descriptives .....	26
Question 1 .....	26
Question 2 .....	28
Exploratory Analysis .....	31

Exploratory Factor Analysis Method.....	32
Exploratory Factor Analysis Results .....	32
Confirmatory Factor Analysis Method .....	36
Confirmatory Factor Analysis Results.....	37
Experiment 1 Discussion .....	38
Chapter 5: Experiment 2 .....	39
Method .....	41
Measures .....	41
Assessing Working Memory.....	41
Assessing Semantic Memory .....	41
Assessing Episodic Memory.....	42
Analysis.....	43
Coding of Variables and Composites .....	43
Outliers and Robustness Analysis.....	43
Results.....	44
Final Analysis Sample and Demographics .....	44
Descriptives.....	45
Question 1 .....	45
Question 2 .....	47
Latent Variables Approach and Modeling Judgments.....	48
Confirmatory Factor Analysis Method .....	49
Confirmatory Factor Analysis Results.....	51
Structural Equation Modeling Method .....	54
Structural Equation Modeling Results .....	55
Experiment 2 Discussion .....	61
Chapter 6: Combined Data.....	61
Analysis.....	62
Coding of Variables and Constructs .....	62
Results.....	63
Final Analysis Sample and Demographics .....	63
Descriptives.....	64
Question 1 .....	64
Question 2 .....	65
Assessing Judgment Model Utility .....	68
Cross-Validation and Parameter Estimation Methods.....	68
Cross-Validation and Parameter Estimation Results .....	72
Combined Analysis Discussion .....	74
Chapter 7: End Remarks .....	75
What Have We Learned?.....	75
Implications of the Distribution of Items on Judgments.....	76
Implications of Memory on Judgments Types.....	79
Conclusions.....	80
Appendices.....	82
Appendix A: Experiment 1 Supplement.....	82
Descriptives of Measures and Composites .....	82
Impetus and Justification of Exploratory Analytic Methods .....	82

Common Factor Analysis vs Principal Component Analysis.....	84
Factor Analysis: Exploratory vs Confirmatory.....	84
Extraction Method: Principal Axis Factoring vs Maximum Likelihood .....	85
Rotation Method .....	85
Exploratory Factor Analysis Supplemental Results .....	86
Appendix B: Experiment 2 Supplement .....	89
Descriptives of Measures and Composites .....	89
Appendix C: Combined Data Analysis Supplement.....	91
Descriptives of Measures and Composites .....	91
Cross-Validation and Parameter Estimation Supplements .....	93
References.....	95

This Table of Contents is automatically generated by MS Word, linked to the Heading formats used within the Chapter text.

## List of Tables

Table 1. Rotated Factor Matrix using Principal Axis Factoring and Varimax rotation .....	34
Table 2. Model fit indices of the 2 and 3-factor memory models.....	37
Table 3. Model fit indices of the four memory models from CFAs on experiment 2 data. Indices that meet the fit guidelines are bolded, with the exception of BIC. ....	53
Table 4. SEM fit indices of the final structural models from experiment 2. ....	55
Table 5. Direct effect estimates for the 3 factor, bi-factor, and 2 <sup>nd</sup> order models from experiment 2 data.....	60
Table 6. BFs for Memory abilities predicting each judgment sum type across each set of analyses. BFs for retrieval ability was assessed using a 2-factor memory model..	66
Table 7. BFs for Memory abilities predicting each judgment sum type across each set of analyses. BFs for retrieval ability was assessed using a 3-factor memory model..	67
Table 8. Table of priors placed on model parameters to be estimated. ....	70
Table 9. Descriptives of the distribution of outcome values for within and out of sample prediction on each model.....	73
Table 10. Median effect size estimates with 95% HPD interval from the posterior distribution using a 3-factor memory model.....	73
Table 11. Mean judgments from data combined from experiment 1 and 2 compared to mean judgments from Sprenger and Dougherty (2006). ....	77
Table 12. Descriptives for each of the tasks and judgment variables of interest in Experiment 1.....	82
Table 13. Table of Eigenvalues and the total variance explained from the EFA using experiment 1 data.....	87
Table 14. Factor Matrices of the unrotated and rotated solutions from the EFA using experiment 1 data.....	88
Table 15. Spearman-Brown task reliabilities for measures used in experiment 1.....	88
Table 16. Construct quality indices of construct quality from CFAs conducted in experiment 1.....	88
Table 17. Descriptives for each of the tasks and judgment variables of interest in Experiment 2.....	90
Table 18. Spearman-Brown task reliabilities for measures used in experiment 2.....	90
Table 19. Construct quality indices of construct quality from CFAs conducted in experiment 2.....	90
Table 20. Descriptives for each of the tasks and judgment variables of interest in the combined data. ....	92

## List of Figures

Figure 1. A comparison of judgment sums by judgment type and distribution. The error bars represent 95% confidence intervals.....	27
Figure 2. The relation between retrieval and working memory ability with judgment sums. The grey area represents the 95% confidence interval region.....	29
Figure 3. Judgment Sums and Memory Factors from the EFA on experiment 1 data. The grey area represents the 95% confidence interval region. ....	34
Figure 4. Standardized results from the CFA conducted on the 2-factor memory model (panel A) and the 3-factor memory model (panel B).....	37
Figure 5. The effect of distribution on judgment sums by judgment type. The error bars represent 95% confidence intervals.....	45
Figure 6. The relation between memory ability and judgment sums as a function of judgment type. The grey area represents the 95% confidence interval region. ....	47
Figure 7. Standardized results from the CFA conducted on the 2-factor memory model (panel A), the 3-factor memory model (panel B), the bi-factor memory model (panel C), and the 2 <sup>nd</sup> order memory model (panel D). ....	53
Figure 8. Structural path models with standardized results for the 3 factor (B), bi-factor (C), and 2 <sup>nd</sup> order (D) models.....	59
Figure 9. The effect of distribution on judgment sums by judgment type. The error bars represent 95% confidence intervals.....	64
Figure 10. The relation between memory ability and judgment sums as a function of judgment type. The grey area represents the 95% confidence interval region. ....	66
Figure 11. Graphical representations of the cognitive models for overall judgments (A) and for each judgment type (B).....	69
Figure 12. The distribution of the observed and predicted values from the overall, probability, and frequency judgment models by prediction type. The dashed lines represent the 95% HPD interval. ....	72
Figure 13. The relation between probability judgments and each memory measure. The grey area represents the 95% confidence interval region. ....	83
Figure 14. The relation between frequency judgments and each memory measure. The grey area represents the 95% confidence interval region. ....	83
Figure 15. Scree plot of the number of components to be extracted from the EFA using Experiment 1 data.....	86
Figure 16. Results from the MAP from the EFA using experiment 1 data. ....	87

# Chapter 1: Grounding Judgment Phenomena in Memory

## Probability Judgments and Subadditivity

One common finding in the judgment and decision making literature is that people often overestimate probability judgments, and one form of this overestimation is known as subadditivity. Subadditivity is the tendency for the probability of an event to be rated as less than the sum of its implicit or unpacked parts. For example, imagine that you received a basket of food that contained fruits, vegetables, and desserts and are asked, “*What is the probability that when you reach into that basket, that you will grab a piece of fruit?*”<sup>1</sup> Let  $P$  stand for your subjective probability,  $F$  stands for fruit,  $V$  stands for vegetables, and  $D$  stand for desserts. The sum of your subjective probabilities,  $P(F) + P(V) + P(D)$ , should total to 100. However, research has shown that subjective probabilities usually sum to be much greater than 100, making the objective probability subadditive to one’s subjective probability.

Tversky and Kahneman were one of the first to call attention to this phenomena and argued that these systematic errors in the judged probability of an event were a result of the explicitness of its description (Tversky & Kahneman, 1983). Tversky and Koehler proposed a theoretical framework, Support Theory, to explain how these different descriptions of the same event can give rise to the biases seen in judged probability (Tversky & Koehler, 1994). In order to test this theory, they designed a series of experiments where they had participants evaluate the probability of an event, given a scenario. For example, participants were told that

---

<sup>1</sup> It is assumed that each food type is mutually exclusive from the others.

nearly 2 million people from the United States die from different causes. They were then asked to estimate the probability of death for various causes. Finding substantial subadditivity in probability judgments, they argued that this bias was due to the judgments being attached not to the event itself, but rather, to the *description* attached to the event. Specifically, they believed that the more descriptive or detailed the event, the more its perceived likelihood increases (Tversky & Koehler, 1994). In other words, when participants estimate the probabilities of causes, the magnitude of their subadditivity depends on the explicitness of the scenario. This is because of the lack of consideration of alternative or unavailable events is the fundamental issue of probability judgment biases. Thus, if one wished to make probability judgments that did not violate normative theory, one needs to fully unpack all alternative events.

The notion that the unpacking of alternative events is critical to the evaluation of the focal event's likelihood reveals a critical assumption. This assumption is that probability judgments are made by comparing the strength of a focal hypothesis<sup>2</sup> with the strength of a set of alternatives. In other words, probability judgments are assumed to engage the judge in a comparison process (Dougherty, Gettys, & Ogden, 1999; Tversky & Koehler, 1994). To make this more concrete, we can go back to the produce basket example. The probability of grabbing a piece of fruit ( $P(F)$ ) is your current focal hypothesis or the likelihood you are assessing. The alternative hypotheses, or states of the world, would be the probability of grabbing vegetables and desserts, ( $P(V) + P(D)$ ). In order to make a probability judgment, you would

---

<sup>2</sup> Focal hypothesis is the current hypothesis or option being considered for judgment.

need to compare the evidence of grabbing a piece of fruit to the evidence of grabbing a vegetable and dessert. That is, you are comparing the likelihood of one state of the world to the likelihood of the other states of the world. This is in contrast to frequency judgments (“*How many pieces of fruit are in the produce basket?*”) which do not involve a comparison process, but rather, are judged by the familiarity or strength of the focal event, in this case, fruit, only (Hintzman, 1984, 1988). Thus, the estimate that you give to your focal hypothesis depends on the degree to which you consider your alternatives.

### *Judgments and Hypothesis Generation*

The work that was done by Tversky and Koehler, in conjunction with the notion that probability judgments involve a comparison process, have a couple of implications: 1) that the generation of alternatives plays a critical role to the assessment of the focal event, and 2) if one wished to understand the systematic errors in probability judgments, one needs to understand the process of how the generation of alternatives occurs. This is exactly what Gettys and Fisher sought out to do in one of their studies (Gettys & Fisher, 1979). They designed an experiment where participants were asked to generate a set of plausible hypotheses for a series of data that were presented one datum at a time. For example, participants were presented with tools used by a plumber and asked to come up with plausible occupations on the basis of these tools. When the first tool was presented, participants gave a set of plausible occupations and were asked to rate the plausibility of each occupation. For each trial, after a new tool was presented, they were asked to re-estimate the plausibility of the set of hypotheses given in the previous trial and

then allowed to generate new hypotheses and drop old ones. The presentation of the data was manipulated such that the first three were consistent with an alternative hypothesis, and the last three data points were chosen to be less consistent with the alternative hypothesis. This was done in order to encourage the generation of new hypotheses. In other words, the experiment was designed such that the hypothesis consistency of the data decreased the plausibility of the initial hypothesis set.

What Gettys and Fisher found was that the active generation of hypotheses was greatly influenced by the subjective assigned probability of the current hypothesis set being considered. Specifically, new hypotheses were triggered when the plausibility of the current hypothesis set decreased. In light of the new data presented in trial N, participants were more likely to generate a new hypothesis if the assessed plausibility of the hypothesis set given on N-1 trial decreased. Further, Gettys and Fisher were interested in understanding the threshold of which triggered the addition of a new hypothesis. Examining the data from trials 2-6, they found that new hypotheses were added if they were at least half as likely as the best hypothesis. Thus, participants appeared to use plausibility estimates in the hypothesis generation process.

This leads to an interesting question. Given a set of data, why would participants not consider all possible alternatives as opposed to employing some sort of plausibility heuristic? Gettys and Fisher believed that the process of hypothesis generation could be modeled as a highly specific recursive memory search (Gettys & Fisher, 1979). In searching for some other possible factors that could play into the hypothesis generation process, they proposed that participants could only entertain a

certain number of hypotheses at a time and that the number of hypotheses that one could consider could be constrained by the individual's own memory limitations. Specifically, the data seemed to suggest that as the number of hypotheses in the current hypothesis set approaches the individual's memory limitations, the quality or probability of the new hypothesis to be included was much stricter. This is crucial because the implications are that the number of alternatives and the likelihoods associated with them can be a function of one's own memory capacity.

### *The Role of Memory in Hypothesis Generation*

More recently, Thomas et al. proposed HyGene, a model of hypothesis generation and probability judgment rooted in memory (Thomas, Dougherty, Sprenger, & Harbison, 2008). This model demonstrates that several systematic errors and phenomena in the decision making literature, including subadditivity, could be accounted for by assuming that basic memory processes are what constrains hypothesis generation. HyGene describes how data extracted from the environment are used to generate hypotheses from memory and how these hypotheses are used to make probability judgments and frame subsequent information search. In accordance with Gettys and Fishers' work, the overall idea of HyGene is that a limited working memory capacity system would lead to an impoverished set of hypothesis. In turn, the impoverished set of hypotheses manifests themselves into systematic errors in probability judgment and subsequent information search.

In HyGene, the subadditivity phenomenon is made explicit through the model. When making a probability judgment, the set of alternative hypotheses one is able to actively consider is impoverished because it is constrained by one's WM capacity.

Several findings purportedly support this mechanism. For example, individual differences in WM capacity were positively correlated with the number of alternative hypotheses one was able to consider (Dougherty & Hunter, 2003a). Additionally, the degree of subadditivity has been shown to be negatively correlated with performance on working memory (WM) tasks. Because WM capacity is linked to the ability to maintain task-relevant information in the focus of attention, these findings suggest that the number of alternative hypotheses that one is able to actively consider is constrained by the limits of their WM. Thus, those with higher WM capacity are able to consider more alternatives than those with low capacity. What follows is that individuals who perform well on WM tasks are less subadditive. This is because they are less likely to overestimate the likelihood of the focal hypothesis when the set of comparisons (alternatives) are more complete (Dougherty & Hunter, 2003a, 2003b; Sprenger et al., 2011; Sprenger & Dougherty, 2006).

These findings notwithstanding, it is important to note that processes of WM and retrieval are intertwined. While the focus of HyGene's explanation of biases in probability judgment has been on WM, what is brought into WM must necessarily be retrieved from LTM. Thus, findings showing correlations between judgment and WM may occur because of individual differences in retrieval, not because of one's WM capacity. Additionally, some work shows that LTM retrieval is important for determining the degree of subadditivity in probability judgment (Sprenger et al., 2011). In a series of experiments, Sprenger et al (2011) investigated the impact of divided attention (DA) during encoding and at retrieval on probability judgments. Because disturbances during the encoding processes can impact what gets stored into

LTM, and thus, one's recall performance, Sprenger et al predicted that disturbances during encoding would lead to one generating less alternative hypotheses. Thus, disturbances during encoding should also lead to increases in judgment magnitude. In order to investigate this, they designed an experiment where participants were placed in either a high or low cognitive load condition and attention was divided during encoding. As predicted, they found that those in the high cognitive load condition recalled fewer alternative hypotheses when compared to those in the low cognitive load group. Further, the overall magnitude of their probability judgments was larger. In a separate experiment where attention was divided at retrieval at four different cognitive load levels, they still found a positive correlation between cognitive load and judgment magnitude. However, the effect of DA on probability judgments during retrieval was much smaller than the effect of DA during encoding. While much of the prior focus has been on the limitations of WM, these findings suggest that errors and biases in the initial storage of information can cascade into errors and biases in judged probability by degrading the retrievability of learned alternatives.

## Chapter 2: Current Study

For probability judgments, the focus has been on how one's WM capacity limits the number of hypotheses one can focus on at any given point in time. In contrast, frequency judgments seem to rely on heuristics such as availability and representativeness. The overarching goal of the current study is to gain insights as to whether the ability to retrieve from long-term memory meaningfully accounts for the

variation in probability judgments. The study will assess individuals' performance in measures of WM and LTM in order to examine its relation variations in probability and frequency judgments. In order to test this, participants will learn a distribution of the to-be-judged items and be asked to make probability judgments on half of the items and frequency judgments on the other half. Participants will then complete a battery of memory tasks that measure the constructs of WM capacity and retrieval from LTM.

Additionally, a study design that allows for contrasting expectations between the two judgments types will be used in order to compare them. Recall that a driving difference in the cognitive processes for each judgment types is that probability judgments involve a comparison process, whereas frequency judgments do not. Taking advantage of that notion, it should not be surprising that the distribution of the strength of the items to which the focal item is compared, would influence its likelihood estimate for probability judgments. This finding, known as the alternative-outcomes effect, is exactly what Windschitl et al found. The distribution of the strength of the alternatives influenced the perceived strength of the focal hypothesis (Windschitl, 2002; Windschitl & Wells, 1998). Specifically, within probability judgments, the magnitude of probability judgments on items are evenly distributed tend to be higher when compared to the judgments on items that are unevenly distributed (Windschitl & Wells, 1998). Theoretically, this paradigm is suitable for the present study because in the context of memory, not considering an alternative item that comes from an evenly distributed set of items should lead to a more

impoverished set of alternatives' strengths, and thus, the consequent is a greater overestimation of the focal item.

### Questions and Hypotheses

Central to this study are the following core questions and hypotheses.

*Question 1: Is there evidence of the comparison process for probability judgments and lack of one for frequency judgments?*

*Hypothesis 1.* First, in regards to judgment type, we expect to replicate the alternative-outcomes effect for probability judgments. Specifically, that the overall magnitude of the sum of probability judgments for items that are evenly distributed should be higher than the overall magnitude of the sum of probability judgments for items that are unevenly distributed. However, we do not expect this relationship to hold when it comes to frequency judgments because they do not involve a comparison process. Thus, the distribution of the alternatives should not play a role in frequency estimation.

*Question 2: What is the relation between each judgment type and memory ability?*

*Hypothesis 2.* In regards to probability judgments, we hypothesize that those who perform better on measures of retrieval from LTM will be less subadditive when compared to those who do not perform as well. Further, we expect this relationship to still hold when controlling for individual differences in WM. Thus, for probability judgments, we expect there to be a negative relation with WM and RA.

*Hypothesis 3.* In regards to frequency judgments, we expect a positive relationship between RA and frequency judgments. This is because historically, frequency judgments have been considered as arising from a memory-based process wherein the

participant recollects specific instances (Brown, 1995, 1997) or estimates the frequency of occurrence based on memory factors such as familiarity and availability (Hintzman, 1984, 1988, Tversky & Kahneman, 1973, 1974). Further, we expect to find no relationship between WM and frequency judgments. A prior study found no relation between WM capacity and frequency judgments (Sprenger & Dougherty, 2006). This makes sense theoretically because frequency judgments do not involve a comparison process and thus, there is no reason to believe that their estimates would relate to working memory.

### Experiments

#### Experiment 1

Experiment 1 examined how individual differences in WM and retrieval from LTM (retrieval ability or RA) relate to the variation in both probability and frequency judgments. Within each judgment type, the frequency of presentation for the to-be-judged items was manipulated such that half of the to-be-judged items were evenly distributed and the other half were unevenly distributed. To see how these judgments relate to memory ability, a battery of memory tests were administered. Composites representing both memory abilities were then used as predictors to test the core hypotheses using a model comparison approach. Finally, exploratory factor analyses were conducted to further understand the latent memory structures themselves and their relation to both judgment types.

## Experiment 2

The purpose of experiment 2 was to see if the findings from experiment 1 were replicable. Findings from experiment 1 suggest that a 3-factor memory model may be better when compared to a 2-factor memory model. Thus, additional measures were added in order to develop both WM and RA at the construct level, by parsing RA up into semantic and episodic memory. Four different measurement models of memory are assessed. Finally, in addition to hypotheses testing using Bayesian model comparison, a latent variables modeling approach will be used to provide converging evidence in understanding the relation between different aspects of memory and different types of judgments.

## Combined Data Analysis

The purpose of combined data analysis was to add additional power and to re-evaluate the central hypotheses. Further, the robustness of the findings from both experiments 1 and 2 was evaluated by analyzing the data through different memory models. Finally, model utility was assessed and parameter estimates were obtained.

## Chapter 3: General Data Analysis Methods and Approach

This section discusses the general approach used in evaluating the a priori hypotheses or predictions. Methods for conducting exploratory analyses are reported in the respective study's analysis section. Information specific to the analysis of a particular experiment or data set is mentioned in its respective section of that study.

### Model Comparison

Hypothesis testing was done using a Bayes factor model comparisons. Bayes factors were used in order to evaluate the strength of evidence for or against the alternative model over the null model, which is typically a model with no predictors. Bayes factors can be interpreted as the odds of one model over the other. For example, a Bayes factor of 3 represents a 3 to 1 odds in favor of the alternative model over the null model. Bayes factors are interpreted on a continuum, with support for the null model given by values less than one and values greater than one providing support for the alternative. The degree of support is indexed by the extremity of the BF with greater support for the null given for values that approach zero and greater support for the alternative given by values that approach infinity. Typically, it is generally accepted that BFs above 3 provide evidence for the alternative model and that BFs below 0.33 provide evidence for the null model. BFs between 0.33-3 are generally viewed as inconclusive (Kass & Raftery, 1995). For each model of interest, the directionality of the parameters of interest will be reported by the Pearson's correlation.

Finally, for the current experiment, the null model always included the participant as a random factor because of the within-subjects design. This will allow for the evaluation of a particular model (or the inclusion of a predictor in a model) on the variance of the dependent variable, over and beyond the nuisance factor of individual differences. For models which focus on just one judgment type (probability or frequency), distribution was also controlled for if it was not of interest.

In general, model comparison or selection typically tells one which model best fits the data *relative* to the other models of which it is compared. However, the utility of a model lies in its predictive quality. If a model has high predictive quality, then the new data predicted from the model fitted with the observed data should be similar to one another. This can be conceptualized as the model's ability to predict unobserved data. Because we are interested in the predictive nature of the data-model fit, cross-validation will be performed. This will be done alongside the general recommended model selection and checking practices (Gelman et al., 2014; Gill, 2015; Lee & Wagenmakers, 2013; Levy & Mislavy, 2016; Lynch, 2007).

#### *Latent Variables Approach*

Because the focus of the study is about one's general memory ability and not about individual task level performance, composites were formed to represent these latent constructs. Composites were formed by first transforming each task's outcome variable to a z-score. The z-scores of the intended composites were then added and then divided by the total number of tasks used to measure that construct. If a participant did not complete one of the tasks used in forming the composite, then the individual had a missing data point for the composite and thus, was not included in any analysis that used the construct.

## Chapter 4: Experiment 1

The purpose of experiment 1 was to examine how one's retrieval ability relates to judgments sums, over and beyond working memory.

## Method

### Participants

Students from the University of Maryland were recruited from the university's online research sign-up program and given course credit for participating. The sample size was determined, a priori, to be 128 total. This total was based on similar studies in the literature. Thus, data were collected until each counterbalancing condition met at least 8 participants.

### Design

A 2 x 2 within subjects randomized block factorial design was used. The two factors were the type of judgment (probability vs frequency) and the to-be-judged items' distribution (even vs uneven). The manipulation of distribution was done such that within each judgment type, half of the frequency of presentation of the to-be-judged were evenly distributed (30, 15, 15, 15, 15) and the other half were unevenly distributed (45, 30, 5, 5, 5). The item frequencies were based off prior studies which also used the alternative-outcomes paradigm. Note that the total number of total presentations (trials) for each subcategory was the same (90). Thus, the only difference between the subcategories was the frequency with which each item was presented. The assignment of the to-be-judged items' sub-categories to each distribution type, the order of making each type of judgment (frequency or probability first), and order of which distribution appeared first when making judgments within a judgment type, were all counterbalanced. This resulted in 16 possible counterbalancing conditions. The blocking factor was participants, which allowed for

the controlling of individual differences between subjects. Additionally, the order of the tasks within the memory battery was randomly generated per participant prior to data collection.

#### Procedure

Participants came into the lab and completed the study visit in one 2 hour testing session. During their visit, participants went through exemplar training where they learned the distribution of the to-be-judged items. Following learning, they made probability and frequency estimates on those items. To understand how WM and RA relate to the variation in judgments, participants completed a battery of memory tasks commonly used to measure WM and LTM. Working memory was assessed by using two complex span tasks, operation span and symmetry span. The ability to retrieve from LTM was assessed by three tasks: delayed free recall, category fluency, and experience fluency. Specific details of these measures are described in the [MEASURES](#) section below. Finally, participants completed a memory free recall task on the items they made judgments on at the end of the visit.

#### Measures

All tasks, with the exception of the working memory tasks, were programmed with PsychoPy version 1.82.01. The working memory tasks were obtained from the Attention & Working Memory Lab at the Georgia Institute of Technology (Unsworth, Heitz, Schrock, & Engle, 2005) and were presented in E-Prime. No modifications were made to those tasks.

## Assessing Probability and Frequency Judgments

The judgment task consisted of three phases: a learning phase, a probability judgment phase, and a frequency judgment phase. Prior to each phase, participants practiced with items from a practice category.

*Learning Phase.* The learning phase consisted of 360 trials where the participants learned the distribution of 20 unique items. The 20 items were evenly split between two categories: tools and clothing. These categories were further split into 2 sub-categories, hats and shoes for clothing, and cooking tools and carpentry tools for tools. This resulted in each sub-category containing 5 items, bringing the total number of items to 20. Each category represented a judgment type (frequency or probability) and each sub-category represented a distribution type (even or uneven). Thus, participants had to learn two distribution types (even and uneven) per category (tools and clothing).

At the beginning of the learning phase, participants were presented with the following scenario:

“Now, imagine that you have two strange and wealthy friends who like to send you gifts. One friend always sends gifts of tools, which would either be a cooking tool or a carpentry tool. The other friend always sends gifts of clothing, which would either be a pair of shoes and or a hat. After opening each package, you repack and throw it onto separate piles (cooking tools, carpentry tools, hats, or shoes) in your garage. Just as you practiced, you will see a series of packages, first with a picture indicating whether it is a package containing a cooking tool, carpentry tool, a hat, or pair of shoes. The package

will open so you will see what item is inside, and you will automatically move on to the next package.”

One learning trial (an opening of a package) contained three pieces of information for the participant: the category of the item, the sub-category of the item, and the particular item itself. Each piece of information was presented in a form of an image with an appropriate text label. The presentation of the category contained an image of a generic brown package with the appropriate label (either tools or clothing). Images for sub-categories consisted of a chef’s hat labeled “cooking” for cooking tools, a red toolbox labeled “carpentry” for carpentry tools, an empty shoebox labeled “shoes” for shoes, and a hat rack labeled “hats” for hats. The timing of the presentation was done so that the category information was presented for 100ms in the center top half of the computer screen, followed by the addition of the sub-category and particular package item for 200 ms, in the bottom left and right of the computer screen, respectively.

A surprise recall task was incorporated into the learning phase such that, on average, on 20% of the trials were followed by the surprise recall. The purpose of including this was to encourage participants to pay attention while learning the distributions of the to-be-judged items. During a surprise recall trial, participants were asked to type in either the last subcategory or the last particular item they opened. Approximately half of the recall trials asked about the last subcategory of the item the participant opened, with the other half asking about the last particular item the participant opened. Participants were instructed on how to complete these recall tasks prior to starting the experimental portion of the task.

*Judgment Phase.* After learning the distributions of the 20 items, subjects proceeded to the judgment phase. Participants were reminded that they now had four piles of gifts in their garage (the four sub-categories) and that they were all repackaged. They were to imagine that they now had a need for one of the items, but since they are repackaged, the participant would need to go to a particular pile and randomly select a package.

*Probability Judgments.* For probability judgments, participants were asked, “Of all the different types of [sub-category] you received, what percentage were of [particular item]?” Participants were instructed to use a number between 0 and 100, with 0% meaning there was no chance that the box contains item [particular item], 100% meaning that it was certain that the box contains item [particular item], and 50% is equivalent to the likelihood of a coin flip landing on heads instead of tails.

*Frequency Judgments.* For frequency judgments, participants were asked “How many packages of [particular item] did you receive?”

To make the judgments, participants typed their answers into a text box, pressed the enter key, and then moved onto the next judgment item. Again, the order of which judgment was made first (probability or frequency) was determined by their counterbalancing condition number. Further, the order in which participants made judgments of the particular items within a sub-category (distribution) was randomized per participant so that they could not use deduction to answer judgment items.

#### Assessing Working Memory

Working memory capacity is seen as the ability to maintain task-relevant information in the focus of attention. The measurement of working memory ability is typically

done by asking participants to complete complex span tasks, which require participants to perform two tasks, a processing component and a memory component, at the same time. The memory component contains varying levels of difficulty (set size). Participants first go through and practice each task separately before performing them at the same time. During that phase, the processing task is interweaved with the memory component. The particular type of task is specified below, under each complex span task section. Finally, participants are told that it is important that they obtain at least an 85% accuracy on the processing component throughout the task. The current study used a shortened version of the two complex span tasks. This was due to study visit time constraints and to consider possible participant fatigue. The shortened complex span tasks were developed by Oswald and colleagues (Oswald, Mcabee, Redick, & Hambrick, 2014) and obtained, with permission, from the Attention & Working Memory Lab at the Georgia Institute of Technology (Unsworth et al., 2005). Prior assessment of the reliability of the shortened versions reveals that they contributed similarly to the standard version's ability to measure WM capacity in addition to predicting general fluid-intelligence (Foster et al., 2015).

For the working memory tasks described below, there are typically two types of outcome scores that one could use, a total score and a partial score. The total score requires that the participant recalls each item within a sub-trial correctly in order to count, where the partial score counts each particular recall instance as either correct or incorrect. In this experiment, the participant's final partial scores were used as the dependent variables. This is typically recommended over the total score because it

typically has more variance and consequently, allows for better discrimination between high and low ability participants (Conway, Kane, & Al, 2005).

*Operation Span.* Participants were asked to memorize the letters presented on the screen in the correct order while solving math problems in between the presentation of the letters. The math problem was presented such that the participant would see the equation and once solved, they clicked the mouse and were presented with a number and had to indicate whether this number was the correct solution by selecting TRUE or FALSE. Following their selection, a letter would be presented on the screen. The number of alternating math problems and letters varied per block. After a certain number of alternating math problems and letters were presented, a recall screen appeared and participants were asked to select the letters that were presented in the correct order.

*Symmetry Span.* Participants were asked to memorize the location of the colored square within a matrix, in the correct order while solving symmetry problems. Participants first saw a large matrix of black and white squares. Once participants determined whether the image was symmetric about the vertical axis, they clicked on the mouse. On the following screen, the participant selected YES or NO to whether the image was symmetrical. After answering the question, participants were presented with a 4x4 matrix where one of the squares were colored red. After a certain number of alternating symmetry problems and colored squares, participants saw a screen containing a blank 4x4 matrix asking them to select the squares in the order presented.

## Assessing Retrieval Ability

The assessment of the ability to retrieve from long-term memory was done using fluency tasks and memory recall tasks. Fluency tasks ask participants to generate as many words as they can within a specified amount of time, within a particular category. Performance on fluency tasks is typically used as an index of one's semantic memory. Retrieval tasks, in particular, from LTM, differ from complex span or short-term memory tasks in that there is a set delay that occurs between the learning of the items that are to be retrieved and the recall of those items. These tasks are commonly used in the memory literature.

For this particular experiment, the latent variable of interest is retrieval ability. The differentiation between the nature (semantic or episodic) of retrieval was not of interest here. In other words, the goal was to assess general retrieval ability from LTM, regardless of the nature of the type of the retrieval process.

*Category Fluency.* For category fluency, participants were presented with four categories, one at a time, and given 2 minutes to generate as many words as they could that belonged to that particular category. Words were typed into a text box that was presented on the computer screen. Participants were given a practice trial to ensure they understood how to type in the answers. The order of the experimental categories was randomized for each participant. The categories of the Category Fluency task were standard from the literature (*sports, animals, fruits*). The practice category was *vegetables*. The average total of the unique words retrieved over the three blocks was used as the dependent variable.

*Experience Fluency.* The experience fluency task was created as a variant to the category fluency Task. Because category fluency has been cited as measuring fluency from memory that is more semantic in nature, the experience fluency task was created to draw from memory could be more episodic in nature. Again, this was done because the main interest of experiment 1 was to look at one's general ability to retrieve from LTM. The episodic variant was created by presenting categories to participants that theoretically could be more episodic in nature. The categories were naming *items that typically go into a backpack, universities/colleges, and female first names*. The practice category was naming *types of relatives*. The instructions for this task was the same as the category fluency task, with the exception that participants were allowed to use proper nouns such as names of people and places. Assessment of this variant is discussed in the [Exploratory Analysis](#) section. The average total of the unique words retrieved over the three blocks was used as the dependent variable.

*Delayed Free Recall.* In the delayed free recall task, participants learned a list of 20 words, with each word presented for 2 seconds at a time. After learning the list of the words, they performed a distractor task and then proceeded to recall as many words as they could, in any order, from the list they had just learned. The time participants had to recall words was 2 minutes. In regards to the distractor task, participants were required to determine whether a set of colored squares, presented briefly on the computer screen, matched the current set of colored squares. There were a total of 3 blocks, with each block containing 4 trials. The four trials differed in the number of squares presented (2, 4, 6, and 8). The time for participants to respond was fixed so that the delay was the same across participants. Thus, the delay was approximately

27 seconds for each participant. The entire sequence was repeated 3 times. The average total number of correct words recalled over the three learned lists were used as the dependent variable.

#### Assessing Recall of the To-Be-Judged Items

Finally, after the completion of the memory battery, participants completed the Judgment Recall task last. Participants were informed that they would now be asked to recall the items that they had made judgments on at the beginning of their study visit. During the task, participants were prompted with each of the four categories, one at a time, and then given 45 seconds to recall the items that belonged to that particular category. The total number of correct items recalled per judgment and distribution type were used as dependent variables.

#### Analysis

##### Coding of Variables & Constructs

*Judgments.* Each participant had 10 frequency judgments and 10 probability judgments, with half of each judgment type coming from either an even or uneven distribution. To examine subadditivity, the total sum of the participant's judgments were used, per sub-category. Thus, each participant had a total of 4 judgment sums (frequency even, frequency uneven, probability even, and probability uneven). To compare the two types of judgments, frequency judgments were re-scaled to be on the range of 0-100 by dividing the estimates by the total number of times that item was presented and then multiplying that number by 100.

*Memory Abilities.* WM ability was represented by average the z-scores of the two working memory tasks and RA ability was represented by averaging the z-scores of the two fluency tasks and the delayed free recall task. Additionally, the recommendation for complex span tasks is to only use data from participants who performed at 85% or above. However, the approach here is similar to robustness analysis. For this project, a second dataset was created to only include participants who met performance criteria. WM and RA composite then were re-calculated. Thus, for all analysis with WM or RA as predictors, a comparable model was run using the data from only participants who met criteria on the complex span tasks.

#### Missing Data

Analysis will include data from all participants who have data on the predictors of interest within each hypothesis test. If a participant did not complete a task that was used in forming a composite or latent factor, they were not included in the analysis that has that composite or latent variable in the model.

#### Outliers and Robustness Analysis

For each experiment, a correlation matrix of the variables and their probabilistic moments will be reported in the appendix. Logit transformation was done on proportion data. No other transformations were done on the data. Additionally, in order to ensure that the findings did not rely on extreme values and to evaluate the robustness of the findings, analyses were conducted on a dataset where potential outliers were removed. Outliers were first identified by finding those whose data points lie outside 1.5 the interquartile range per task. These participants were then

removed from the dataset and the composites were re-calculated. Because of the performance criteria for WM tasks mentioned previously, this resulted in each hypothesis being evaluated with 4 variants of the dataset: 1) a full dataset, 2) a dataset with only those who met performance criteria on the WM tasks, 3) a dataset with potential outliers removed, 4) a dataset with potential outliers removed and who only met performance criteria.

#### Software

Descriptives and Bayesian model comparison were conducted in R version 3.4.2. Exploratory factor analyses were conducted using SPSS version 24 and MPLUS version 7.

### Results

#### Final Analysis Sample and Demographics

One-hundred forty-two University of Maryland undergraduates (mean age =  $19.16 \pm 2.68$ , 106 females) completed the study. Out of the 142 participants who completed the study, 6 participants had missing data on one or more of the tasks. Out of those 6 participants, 5 were missing data on a task used to form a composite. However, all analysis will include all participants that have data on the predictors of interest. Thus, the final main analysis will include 137-142 participants (103 females, mean age =  $19.16 \pm 2.68$ ). The total number of participants for the dataset containing only those who met performance criteria on working memory measures was 118. The number of participants in the dataset with potential outliers removed was 129, and the number of

participants in the dataset with potential outliers removed and who also met performance criteria on working memory measures was 112.

#### Descriptives

A correlation matrix of the judgments, composite memory abilities, and task variables are located in the appendix.

#### Question 1

*Is there evidence of the comparison process for probability judgments and lack of one for frequency judgments?*

The primary hypothesis of interest was whether judgment magnitude varied as a function of judgment type and distribution. I hypothesized that the overall magnitude of the sum of judgments for items that are evenly distributed should be higher than the overall magnitude of the sum of probability judgments for items that are unevenly distributed. However, I do not expect the distribution of the alternatives to play a role in frequency judgments because they do not involve a comparison process. Thus, the expectation is that there should be a main effect of judgment type and a meaningful interaction between distribution and judgment type in explaining judgment sums.

Figure 1 shows judgment sums as a function of judgment type and distribution.

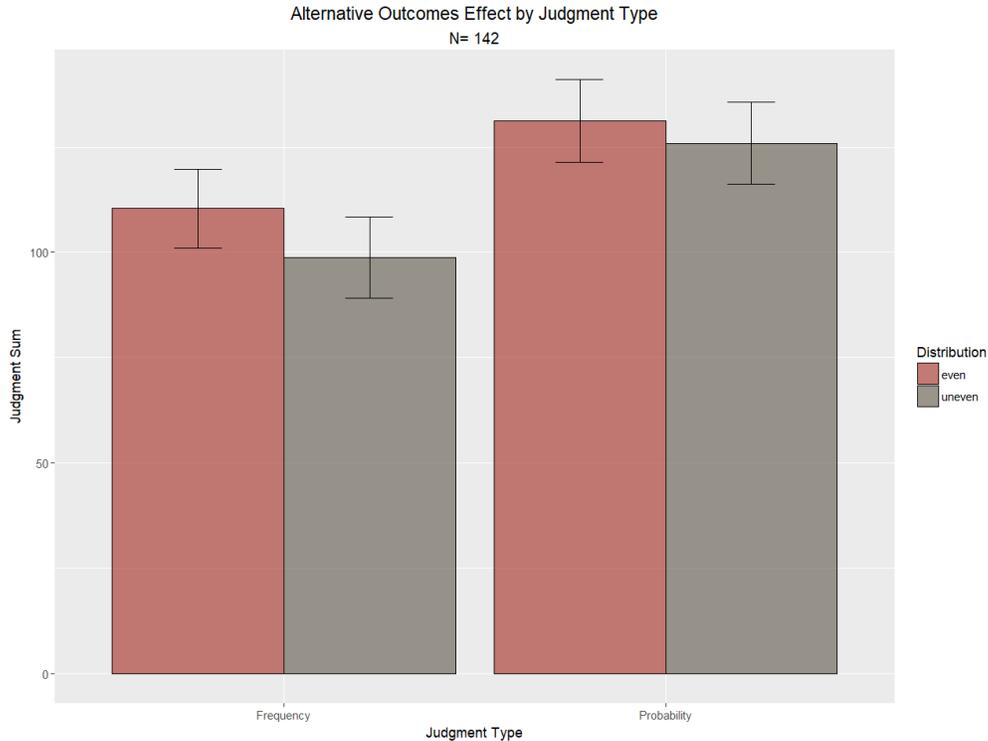


Figure 1. A comparison of judgment sums by judgment type and distribution. The error bars represent 95% confidence intervals.

As can be seen, judgment sums for both probability and frequency are slightly greater for the even distribution compared to the uneven distribution, and overall probability judgments are higher than frequency judgments. In order to test this hypothesis, the following models were run:

$$Judgment \sim Judgment\ Type + Distribution + Judgment\ Type * Distribution$$

$$Probability\ Judgment \sim Distribution$$

$$Frequency\ Judgment \sim Distribution$$

Although both patterns are consistent with prior work (Dougherty & Hunter, 2003a, 2003b; Sprenger et al., 2011; Sprenger & Dougherty, 2006) only the effect of judgment type was supported statistically ( $BF = 11957.43 \pm 11.52\%$ ). The test of the main effect of distribution was inconclusive though slightly favored the null

hypothesis ( $BF = 0.586 \pm 8.14\%$ ). Critically, the test of the interaction yielded modest support for the null hypothesis: Comparing the interaction model with the two-main effects model yielded a  $BF = 0.179 \pm 8.14\%$ , which suggests that the two-main effects model is over 5 times more likely than the interaction model. Moreover, comparison of the interaction model with the main effect model that includes only judgment type provided strong evidence in favor of the one-main effect model over the interaction model. Taken together, these results suggest that neither distribution nor the interaction between distribution and judgment type is useful in explaining judgment sums. Focusing on each judgment type, results were inconclusive on whether the type of distribution was an important predictor in the model for probability judgments ( $BF = 0.387 \pm 1.13\%$ ). However, for frequency judgments, results showed that the data was much more likely to be observed from the alternative model than the null model ( $BF_{10} = 781.914 \pm 1.05\%$ ).

## Question 2

*What is the relation between each judgment type and memory ability?*

The hypotheses of interests here are about the expected relation between each memory ability and judgment sum as a function of judgment type. I hypothesized that RA should have a negative relation to both probability and frequency judgments. For WM, the prediction was that probability judgments would have a negative relation and that there should be no relation with frequency judgments. Figure 2 shows the relation between each judgment type and memory ability. To test these relations, the following models were run:

*Probability Judgment ~ Distribution + RA + WM*

## Frequency Judgment ~ Distribution + RA + WM

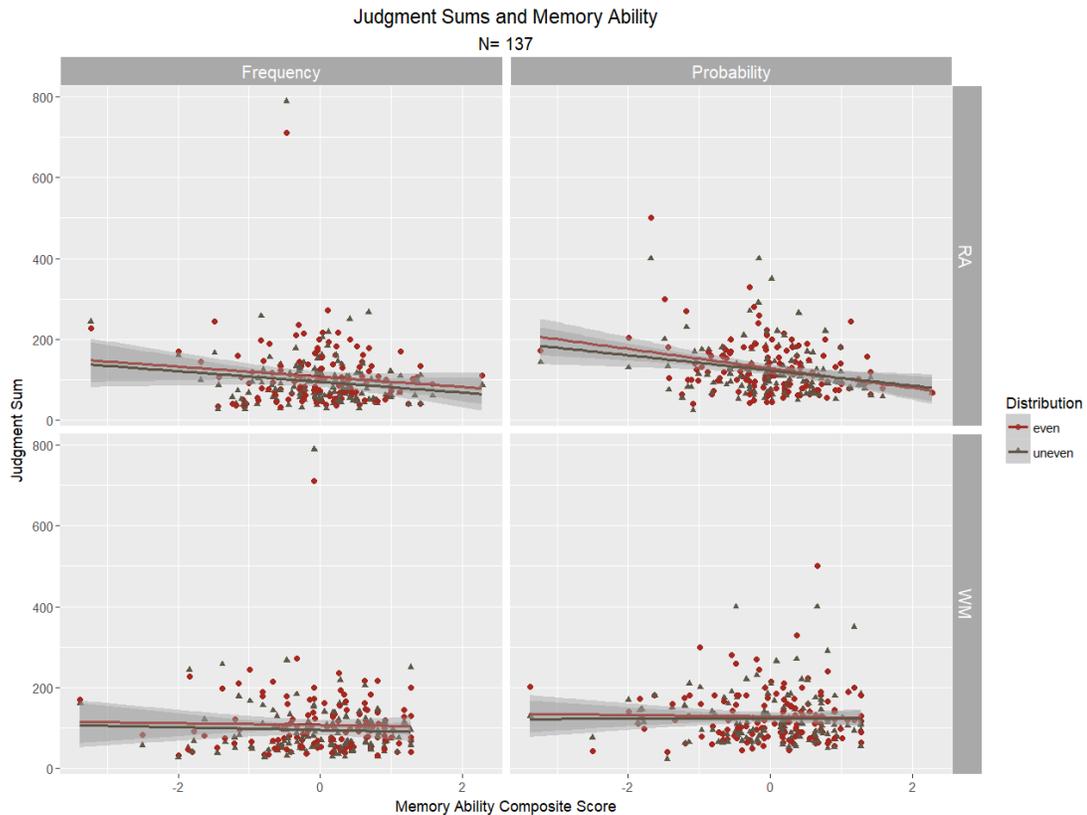


Figure 2. The relation between retrieval and working memory ability with judgment sums. The grey area represents the 95% confidence interval region.

**Probability Judgments.** The plot shows that probability judgment sums are negatively correlated with retrieval ability ( $r = -0.257$ ) suggesting that those who perform better on long-term memory retrieval tasks tend to have lower probability judgment sums. Results from the model comparison revealed support that retrieval ability predicted probability judgments ( $BF = 23.063 \pm 0.89\%$ ) and this held when controlling for individual differences in WM ( $BF = 35.058 \pm 4.89\%$ ). This suggests that the data are more likely to be observed in models that include retrieval ability as a predictor when compared to models without it.

In regards to WM, unlike previous findings, working memory appears to have almost no relation with probability judgment ( $r = -0.011$ ). Additionally, results from

the model comparison revealed no support that a model containing WM improved data-model fit over the null model ( $BF = 0.329 \pm 3.15\%$ ). This shifted to inconclusive evidence when controlling for all other predictors (distribution, RA, and participant) in the model ( $BF = 0.757 \pm 3.9\%$ ).

As mentioned before, each model was run with and without those who performed to criteria in complex span tasks, and each of these sets was run with and without potential outliers. These analyses only changed the magnitude of the BFs. For example, when controlling for the other model predictors, the range of BFs for WM was 0.754 to 1.086. For RA, the range of BFs was 20.501 to 49.752. Thus, the statistical conclusions from the main dataset are robust to the decision to include or exclude influential data points and the performance criteria of the WM span measures. Subsequent hypotheses conclusions will report any changes that may yield a different interpretation of the data.

*Frequency Judgments.* Figure 2 shows a negative relation between frequency judgments and RA ( $r = -0.131$ ), implying that those who perform better on long-term retrieval tasks tend to have lower frequency judgment sums. However, results from model comparison were inconclusive when looking at RA as a predictor of frequency judgments ( $BF = 0.797 \pm 3.22\%$ ), which held even when controlling for WM and distribution ( $BF = 1.225 \pm 6.47\%$ ).

In regards to WM, similar to probability judgments, the plot suggests a very weak to no relation between WM and frequency judgment sums ( $r = -0.032$ ). Model comparison results revealed inconclusive evidence for WM being predictive of

frequency judgment sums ( $BF = 0.469 \pm 4.01\%$ ). This was also the case when controlling for distribution and RA ( $BF = 0.611 \pm 6.4\%$ ).

### Exploratory Analysis

The purpose of the following exploratory factor analyses was to understand the covariation among the observed variables as a function of the assumed latent constructs (Hancock & Mueller, 2010). Analysis conducted using Bayesian model comparison evaluated memory ability using composite scores. These scores came from measures that were assumed to theoretically tap into their respective construct. The idea here is that the tasks used in this experiment are effect indicators of a latent ability. Further, because a newly developed task was used, experience fluency, it is of interest to evaluate whether all measures loaded appropriately onto their latent factor. Thus, two additional types of analysis were conducted, an exploratory factor analysis (EFA) and a confirmatory factor analysis (CFA).

The purpose of the EFA is to assess the latent nature of the memory tasks and to provide additional insights into understanding the variation in judgments. The data were re-analyzed as follows: (1) the latent variables were extracted out of the different memory measures, (2) for each latent variable, a factor score was derived for each participant, and (3) probability and frequency judgment estimates were each regressed onto the factor scores. This allows for possible insights on how individual differences in these factors influence and relate to probability and frequency judgments. Finally, while an EFA starts with the data, the purpose of the CFA is to start with the theoretically derived measurement model to evaluate the assumed factor

structure. That is, the a-priori latent constructs will be assessed by checking if the observed measures load appropriately onto their respective memory construct.

#### Exploratory Factor Analysis Method

Principal Axis Factoring (PAF) was conducted on the 5 measured memory variables in order to extract latent variables. Instead of using just one dependent variable per task, the outcomes of each block within each task was used. Thus, across the 5 tasks, there were a total of 15 variables that went into the factor analysis. The number of factors to be extracted will depend both on theory and interpretability. In this particular scenario, at minimum, it is expected that at least two factors should arise from the data (WM and RA). However, since the retrieval tasks were comprised of fluency and recall measures, a 3-factor solution would be justifiable. To assist with interpretation, orthogonal rotation was used, specifically, the results below are based on varimax rotation. Finally, factor scores were then saved for each participant using the regression method. These scores were then used to predict the different types of judgments. The impetus and further justifications for the specific methods used for this EFA can be found in the appendix.

#### Exploratory Factor Analysis Results

*Rotated Factors Solution.* Principal components analysis was first run on the data to aid in determining the number of factors one should extract. The scree plot and results from the Minimum Average Partial procedure (MAP) suggested that the number of factors to extract was 2 and 3, respectively. In concert with theory, interpretability of the factors, and the overarching goal of the EFA, the final solution was a 3-factor

solution. Additional output from the EFA, including scree plot, results from the MAP procedure, and variance explained of the 2 and 3-factor solution are included in the appendix.

Factor 1 had high loadings of the fluency tasks. The high loadings of factor 2 were with DFR and Symmetry Span. The high loadings on factor 3 were Operation Span. Below is a table of factor loadings. Factor scores were derived then used as predictors in a model that looked at frequency judgments and a model that looked a probability judgment. Bayesian model comparison was used to evaluate the top model for each judgment type, in addition to each factor.

	<i>Factor 1</i>	<i>Factor 2</i>	<i>Factor 3</i>
<i>Category Fluency 1</i>	<b>.679</b>	.368	.035
<i>Category Fluency 2</i>	<b>.765</b>	.264	-.026
<i>Category Fluency 3</i>	<b>.673</b>	.189	.095
<i>Experience Fluency 1</i>	<b>.758</b>	-.084	.083
<i>Experience Fluency 2</i>	<b>.552</b>	-.122	-.032
<i>Experience Fluency 3</i>	<b>.657</b>	-.016	.173
<i>Delayed Free Recall 1</i>	.170	<b>.452</b>	.199
<i>Delayed Free Recall 2</i>	.047	<b>.690</b>	.129
<i>Delayed Free Recall 3</i>	.137	<b>.692</b>	.069
<i>Symmetry Span 3</i>	.028	<b>.460</b>	.342
<i>Symmetry Span 4</i>	-.047	<b>.493</b>	.298
<i>Symmetry Span 5</i>	-.086	<b>.457</b>	.393
<i>Operation Span 4</i>	.061	.387	<b>.459</b>

Operation Span 5	.094	.161	<b>.690</b>
Operation Span 6	.128	.210	<b>.673</b>

Table 1. Rotated Factor Matrix using Principal Axis Factoring and Varimax rotation

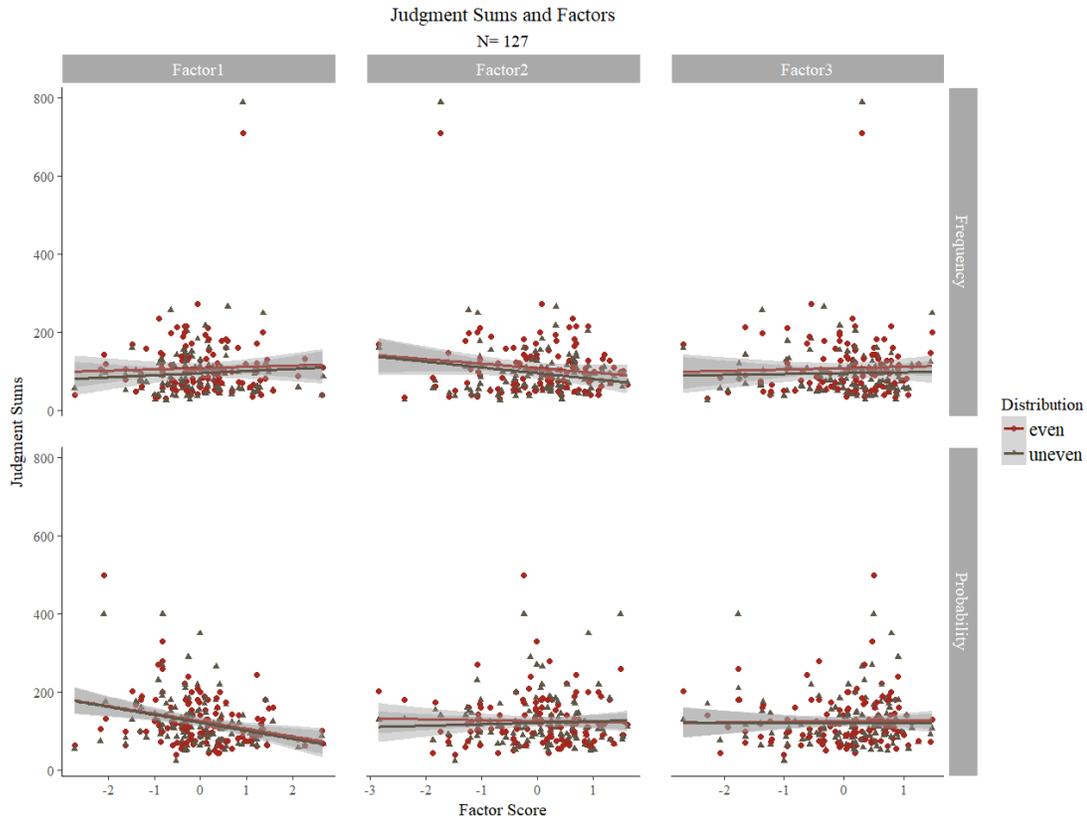


Figure 3. Judgment Sums and Memory Factors from the EFA on experiment 1 data. The grey area represents the 95% confidence interval region.

*Predicting Judgments.* Figure 3 shows the relation between each judgment type and memory factor. Recall that analysis from model comparisons revealed that RA was an important predictor of probability judgment sums and that for frequency judgments, both WM and RA drew inconclusive evidence. The plots in figure 3 suggest that for probability judgments, factor 1, which contained high loading from the fluency tasks, has the strongest relation with judgment sums out of the 3 factors. The remaining factors show either little to no relation with probability judgment sums. For frequency judgments, all of the factors seem to be weakly related, with

factor 2, which contains high loadings of symmetry span and DFR, having the strongest relation relative to the other factors. To test these relations, the following models were run:

*Probability Judgment ~ Distribution + Factor 1 + Factor 2 + Factor 3*

*Frequency Judgment ~ Distribution + Factor 1 + Factor 2 + Factor 3*

*Probability Judgments.* For probability judgments, the top model was a single factor model, Factor 1 (BF = 47.736 ±2.37%). This model was still preferred over the second best fitting model, a 2 factor model with factors 1 and 3 (BF = 3.107 ±3.8%). This suggests that the factor with high loadings from both fluency tasks best predicted the variance in probability judgment sums. When controlling for each other predictor in the model (factor 2, factor 3, distribution, and participant), there was strong support for factor 1 ( $r = -0.287$ , BF = 27.463 ±6.07%). Results were inconclusive for factor 2 ( $r = 0.009$ , BF = 0.399 ±11.96%) and factor 3 ( $r = 0.010$ , BF = 0.444 ±8.06%). Finally, there was no support that distribution should be included in the model ( $r = -0.033$ , BF = 0.320 ±4.4%).

*Frequency Judgments.* The top model for frequency judgments was a 2-factor model which had distribution and factor 2 as predictors (BF = 6549.973 ±26.57%). However, when compared to the second best fitting model, a 3-factor model which added factor 1, results were inconclusive as to which had better predictive fit (BF = 1.989 ±26.82%). Looking at the effect of distribution, with all other predictors controlled for, results showed strong support ( $r = -0.085$ , BF = 6079.628 ±23.16%). When examining each factor on its own, controlling for all other predictors in the model, results were inconclusive. The BF for factor 1 was 0.939 ±22.65% ( $r =$

0.053), the BF for factor 2 was  $2.060 \pm 22.97\%$  ( $r = -0.149$ ), and the BF for factor 3 was  $0.982 \pm 22.86\%$  ( $r = 0.031$ ). In line with the findings thus far, this suggests that the best predictor, relative to all others in the model, is distribution.

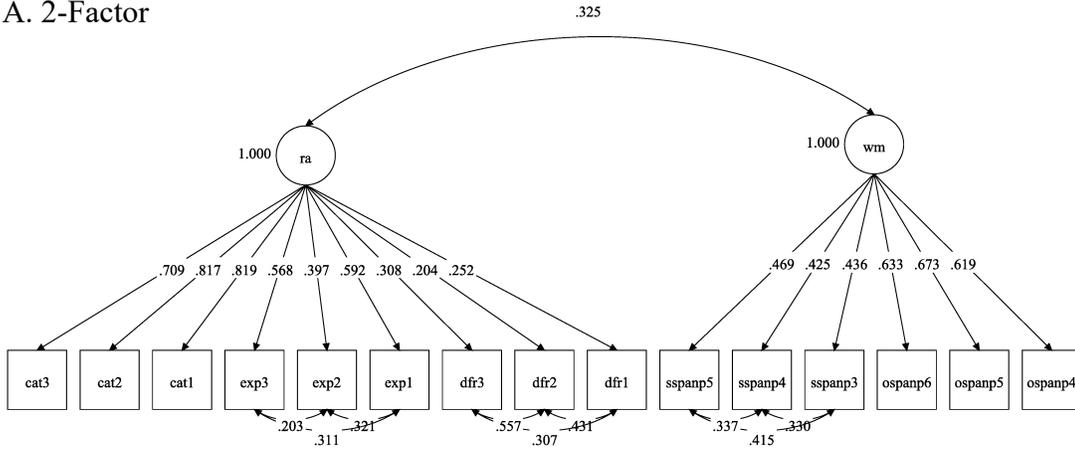
#### Confirmatory Factor Analysis Method

An interesting finding from the EFA results was the split between the working memory tasks between factors 2 and 3 and the retrieval tasks between factors 1 and 2. As mentioned previously, one way of assessing the assumptions of the measures and their respective latent constructs is to conduct a CFA. CFAs can be thought of as a special case of EFA, in which CFAs impose more restrictions on a measurement model, with these restrictions usually being based on a-priori theoretical assumptions.

The indicator variables used in the following CFA were the same as what was used in the EFA. Additionally, the error variances of indicator variables were allowed to co-vary if they were within the same task and if the task structure was highly similar. Latent factors were also allowed to co-vary. Finally, the Satorra-Bentler Scaling method was used to adjust for non-normality (Satorra & Bentler, 2010). Two measurement models were tested, a 2-factor memory model and a 3-factor memory model. For the 2-factor memory model, the latent variables were WM and RA, similar to how the composites were formed in the prior model comparisons. The 3-factor memory model was similar to the 2-factor model, with the exception that RA was parsed into semantic memory and episodic memory. This was guided by the scatterplots of retrieval ability at the task level and the grouping of the factors from the EFA.

## Confirmatory Factor Analysis Results

### A. 2-Factor



### B. 3-Factor

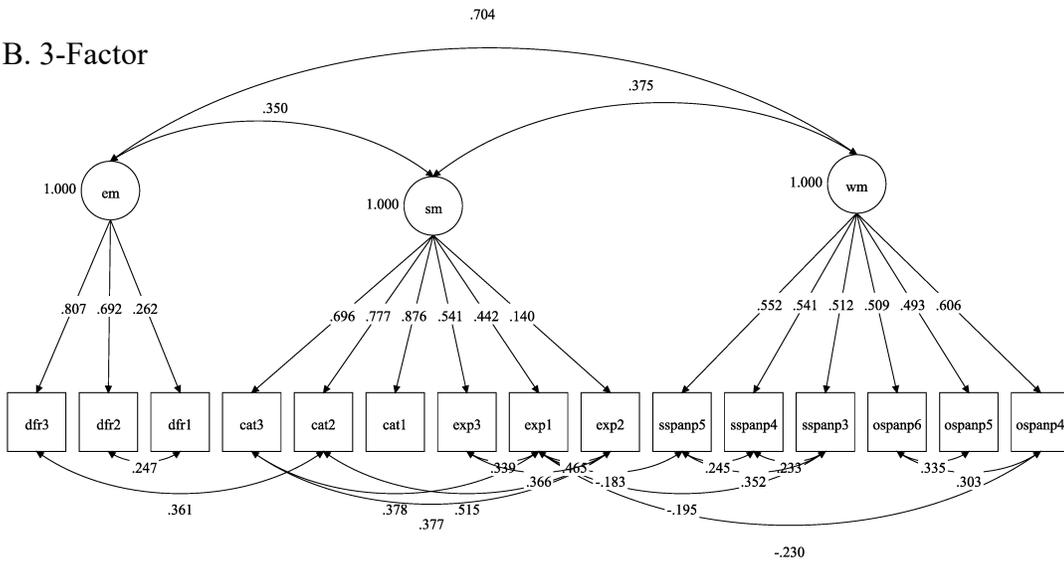


Figure 4. Standardized results from the CFA conducted on the 2-factor memory model (panel A) and the 3-factor memory model (panel B).

	RMSEA with 95%			
	CI	CFI	SRMR	BIC
Model A - 2 Factor	0.072 (0.049, 0.094)	0.914	0.103	9947.518
Model B - 3 Factor	0.048 (0.01, 0.074)	0.962	0.059	9915.205

Table 2. Model fit indices of the 2 and 3-factor memory models.

Task reliabilities are reported in the appendix. Figure 4 shows the 2 memory measurement models and the resulting standardized path results. Table 1 shows

model fit indices from the CFA. Using the measurement fit guideless proposed by Hu & Bentler (Hu & Bentler, 1995), only the 3-factor model, had good fit when compared to the 2-factor model. Additionally, because the 2-factor model is nested within the 3-factor model, a chi-square difference test can be used to determine whether the addition of more parameters was worth the increase in overall data-model fit. Results revealed that the loss of degrees of freedom was worth the increase in fit,  $\chi^2(22) = 67.1941$ ,  $p = 6.474471e-07$ . Additionally, the BIC values were compared and results also revealed strong evidence for the 3 factor model over the 2-factor model (BF = 10391513). Construct quality indices of the latent factors for each model are reported in the appendix.

### Experiment 1 Discussion

Model comparison results suggest that overall, the type of judgment (probability of frequency) was meaningful in predicting the variance in judgment sums and that the frequency of which an item is presented has predictive value to the estimate of the particular item. For probability judgments, the best overall model was either a one or 2-factor model which included RA and WM as predictors. In regards to memory abilities, there was support that retrieval ability plays a role in accounting for the variance in probability judgments. For frequency judgments, the top models decisively included distribution.

Additionally, exploratory factor analysis suggests that different aspect of retrieval ability may differentially relate to each judgment type. Specifically, the fluency tasks and the delayed free recall tasks may be tapping into different

constructs (semantic memory and episodic memory, respectively) and that these constructs may relate differently to the different types of judgments.

## Chapter 5: Experiment 2

The purpose of experiment 2 was to follow up on the findings in experiment 1. First, we wanted to test the replicability of experiment 1 results. Further, because we are interested in individual differences in latent abilities, more measures were included in experiment 2 in order to have more refined latent constructs and analyses. In experiment 1, retrieval from long-term memory was measured as one general latent ability (retrieval ability) and Bayesian model comparison revealed the data are more likely to be observed under models that included retrieval ability as a predictor when compared to models without it. However, this assumes that retrieval from long-term memory has the same relation or nature with each judgment type across the types of long-term memory being accessed. The exploratory analysis conducted in experiment 1 seems to suggest that this may not be the case. That is, retrieval from long-term memory and different judgment types may vary depending on the nature of the type of retrieval.

To better assess this, experiment 2, adds more measures such that retrieval from long-term memory was split into retrieval from memory that was more semantic in nature (semantic fluency or semantic memory), and retrieval from memory that was more episodic in nature (episodic memory). The parsing of long-term memory retrieval into semantic and episodic memory was guided by the findings from both the EFA and CFA in experiment 1.

Hypothesis testing in experiment 1 was conducted by using composite variables to represent unobserved latent abilities. However, the use of composite variables often carry measurement error and can attenuate the relation between it and the outcome of interest. The inclusion of more indicator (measured) variables in experiment 2 is to take a more intentional latent modeling approach. Analysis of the central core hypotheses will still be assessed using Bayesian model comparison to evaluate the strength of evidence for each memory ability. To provide converging evidence, additional analysis using a latent variables approach will be conducted by use of structural equation modeling (SEM). An advantage of using a latent variables approach is that it allows for the handling and accounting of measurement error that may attenuate relations between the latent variables of interest and a particular outcome variable.

Finally, the testing of the central questions and hypotheses in experiment 2 are conducted in a similar fashion to those in experiment 1. However, instead of modeling judgments as a function of a 2-factor memory model (WM and RA), judgments were modeled using a 3-factor memory model, WM, semantic memory (SM), and episodic memory (EM). Because we are interested in the unique contribution of each latent memory ability, a higher sample size to gain additional power together with more indicator variables also allows us to assess the memory measurement model to assure that the measures load onto their respective latent abilities appropriately. Altogether, experiment 2 will provide more insights into how memory, and the way it is measured, relate to judgments as a function of judgment type and the type of memory being evaluated.

## Method

The methods used in experiment 2 were generally the same as the methods outlined in experiment 1, with the exception that participants completed more memory tasks, which will be described below in the [Measures](#) section. Any other differences in the experiment design or analysis will be mentioned in the relevant section.

## Measures

All of the measures that were included in experiment 1 were included in experiment 2. Thus, only new tasks or task changes will be discussed in this section. Tasks were programmed with PsychoPy's version 1.82.01 or E-Prime 2.0 build 2.0.10.353.

### Assessing Working Memory

*Reading Span.* Similar to operation and symmetry span used in study 1, the shortened version of the reading span was included in experiment 2. For this task, participants were asked to make true false judgments of sentences on the screen between the presentations of the letters. Similar to the other complex span tasks, the number of alternating sentence comprehension problems and letters varied per block, and after a certain number of alternating sentences and letters were presented, a recall screen appeared and participants were asked to select the letters that were presented in the correct order.

### Assessing Semantic Memory

The assessment of semantic memory was done by using the category and experience fluency tasks, which were described in experiment 1. An additional fluency task was added and is described below.

*Letter Fluency.* The letter fluency task is similar to the category fluency task in that participants are given 2 minutes to generate as many words as they can that begin with a specified letter. Again, the order of the experimental letters was randomized for each participant. The letters used in the task were standard from the literature (*A, S, F*). The practice letter was *L*.

#### Assessing Episodic Memory

Episodic memory was measured by using the delayed-free recall task from experiment 1 and the three additional measures. Each episodic memory task included a learning phase and then a recall phase which are described below.

*Cued Recall.* The cued recall task requires participants to go through a learning phase and then a retrieval phase. During the learning phase, participants are presented with 40 word pairs, each for 6 seconds. After the learning phase, participants go into a self-paced retrieval phase where they were cued by with the first word of a word pair and then asked to retrieve the second word of the word pair. An individual's score was the total number of words correctly recalled.

*Picture-Source Recognition.* Participants were informed that they would be presented with images, each appearing on one of four locations on the screen, and to try to correctly recall the location of which the image appeared. During the learning phase, each image was presented for 1 second in one of the four locations on the screen. After the learning phase, participants moved onto the retrieval phase in which 30 old and 30 new pictures were presented. While these images were presented, participants had to respond with which of the four locations the image appeared in or whether the picture was new. They were given 5 seconds to respond. Participants had two scores,

in order to differentiate between source accuracy and correct rejection. In previous studies, the score on this task has simply been the proportion of correct responses.

The two scores were the proportion of old items that were correctly identified in the correct quadrant and the proportion of new items correctly identified as new.

*Gender-Source Recognition.* In this task, participants were told that they would hear 30 words, read in either in a male or female voice and to try to remember the word and the gender of the voice. During the retrieval phase, 30 old and 30 new words were presented on the computer screen and participants were given 5 seconds to respond with whether the word was new or old, and if old, whether the gender of the voice that the word was read in was a male or female voice. Similar to Picture-Source Recognition, participants had two scores, source accuracy and correct rejection. The two scores were the proportion of old items that were correctly identified with the correct gender and the proportion of new items correctly identified as new.

### Analysis

#### Coding of Variables and Composites

*Memory Abilities.* Composites were formed in a similar manner to experiment 1 in that for each construct, the outcome variable for each task was then z-scored and averaged to form each memory ability composite.

#### Outliers and Robustness Analysis

Outliers and tasks with a performance threshold were also handled in a similar manner to experiment 1. Thus, each analysis was conducted on 4 datasets: one that

included all data points, one that included only those who met performance criteria on all complex span tasks, one with potential outliers removed, and one with potential outliers removed and contained only those who met performance criteria on all complex span tasks. Again, for each dataset, composites were re-calculated such that the composite score only carried information of those who were included in the analysis.

### Results

#### Final Analysis Sample and Demographics

Two-hundred and forty (mean age =  $19.06 \pm 1.49$ , 161 females) University of Maryland undergraduates completed the study. Out of the 240 participants who completed the study, 10 participants had missing data on one or more of the tasks. Of those 10 participants, 3 did not complete the judgment task and 4 did not complete any of the memory measures correctly. Those 7 participants were excluded from the final analysis dataset. The remaining 3 participants, and in general, all participants, were included in any analysis contained variables that they had data for. Thus, the final main analysis will include 233 participants (156 females, mean age =  $19.05 \pm 1.50$ ). The total number of participants for the dataset containing only those who met performance criteria on working memory measures was 188. The number of participants in the dataset with potential outliers removed was 197, and the number of participants in the dataset with potential outliers removed and who also met performance criteria on working memory measures was 173.

## Descriptives

A correlation matrix of the memory composites, task variables, and judgment sums are provided in the appendix.

### Question 1

*Is there evidence of the comparison process for probability judgments and lack of one for frequency judgments?*

The hypothesis was that the overall magnitude of probability judgment sums for evenly distributed items would be greater when compared to the magnitude of unevenly distributed. Further, no meaningful difference was expected for frequency judgment sums as a function of distribution.

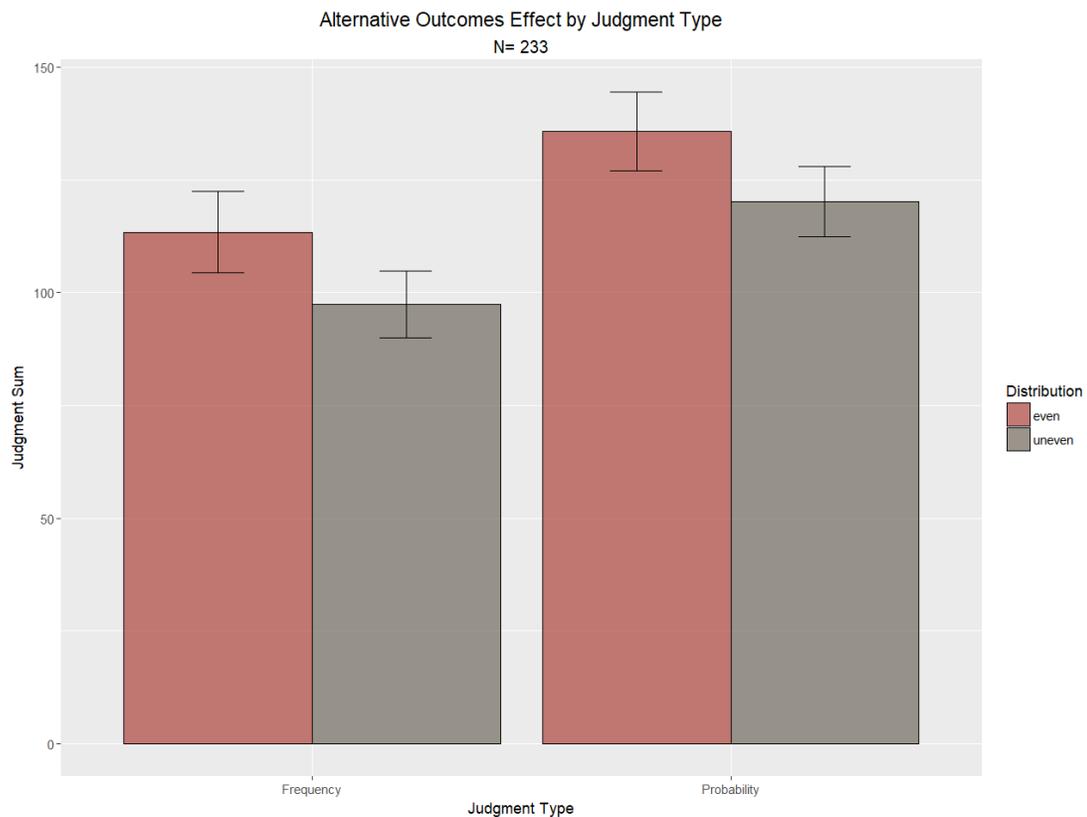


Figure 5. The effect of distribution on judgment sums by judgment type. The error bars represent 95% confidence intervals.

Figure 5 shows judgment sums as a function of judgment type and distribution. Somewhat similar to experiment 1, the overall judgment sums for items that were evenly distributed are greater when compared to those from the uneven distribution. In experiment 1, results revealed that judgment magnitude as a function of distribution was inconclusive for probability judgments but decisive for frequency judgments. The same models were used to test this prediction for experiment 2:

*Judgment ~ Judgment Type + Distribution + Judgment Type \* Distribution*

*Probability Judgment ~ Distribution*

*Frequency Judgment ~ Distribution*

Similar to experiment 1, there was decisive support of the effect of judgment type, when controlling for all other terms in the model ( $r = 0.147$ ,  $BF = 123643.7 \pm 6.73\%$ ). Additionally, the test of the interaction model with the two-main effects model revealed that the data are over 8 times more likely ( $BF = 0.121 \pm 3.36\%$ ) to be observed under the 2 main-effect model over the interaction model. Finally, in contrast to experiment 1, there was support that for distribution as a predictor in modeling the judgment sums ( $r = -0.103$ ,  $BF = 106.689 \pm 3.32\%$ ).

Focusing now on each judgment type, for frequency judgments, similar to the findings in experiment 1, results also showed strong support for the effect of distribution ( $BF = 19978233 \pm 0.76\%$ ). For probability judgments, there was also strong support for the effect of distribution on probability judgment sums ( $BF = 10542807 \pm 0.66\%$ ). Particularly, that the sums from the evenly distributed items were greater than those in the unevenly distributed items. Overall, experiment 2 replicated the alternative outcomes effect for probability judgments.

## Question 2

*What is the relation between each judgment type and memory ability?*

The findings from experiment 1 suggest that those who perform better on long-term memory retrieval measures have lower judgment sums when compared to those who do not perform as well. Further, exploratory analyses suggest that semantic fluency measures may be the driving force behind this finding. For the current experiment, retrieval ability was parsed into semantic memory (SM) and episodic memory (EM). The prediction is that RA would still be important to modeling probability judgment, but that SM would have the strongest support among the retrieval abilities. Figure 6 shows the relation between probability judgment sums and each memory ability.

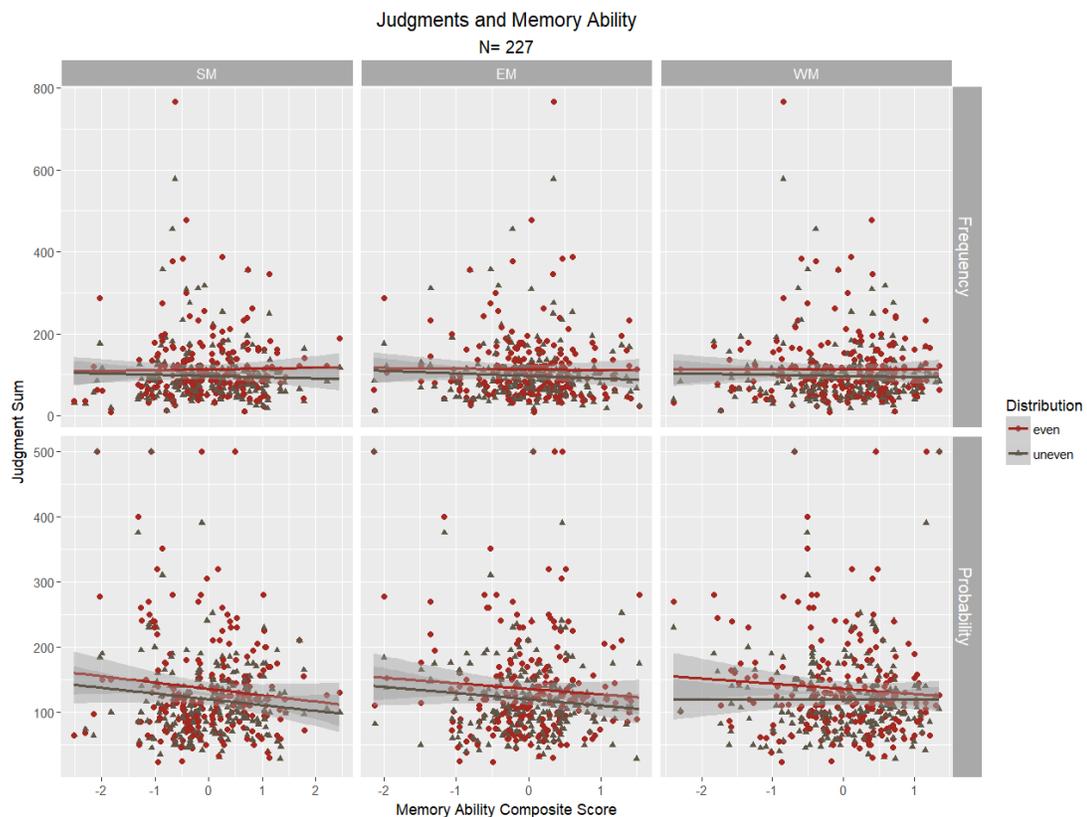


Figure 6. The relation between memory ability and judgment sums as a function of judgment type. The grey area represents the 95% confidence interval region.

The scatterplot reveals, there's a negative correlation between each memory ability and probability judgment sums. The strongest correlation is between probability judgment sums and semantic ability ( $r = -0.102$ ), followed by episodic memory ( $r = -0.079$ ), and then working memory ( $r = -0.038$ ). For frequency judgments, the strongest correlation was with retrieval from episodic memory ( $r = -0.034$ ), followed by working memory ( $r = -0.013$ ), and then by semantic memory ( $r = -0.006$ ). However, the correlations among all judgment types and abilities were generally weak. To test the relations of each memory ability and judgment sums, the following models were evaluated:

$$\text{Probability Judgment} \sim \text{Distribution} + SM + EM + WM$$

$$\text{Frequency Judgment} \sim \text{Distribution} + SM + EM + WM$$

Similar to experiment 1, results were inconclusive for WM as a predictor in both probability ( $BF = 0.554 \pm 7.25\%$ ) and frequency judgments ( $BF = 0.599 \pm 7.77\%$ ). Further, results revealed inconclusive evidence of either SM or EM being predictive of probability judgments ( $BF = 0.923 \pm 7.35\%$  and  $BF = 0.562 \pm 7.26\%$ , respectively). This was also the case for frequency judgments (SM:  $BF = 0.562 \pm 7.72\%$ , EM:  $BF = 0.558 \pm 7.55\%$ ), which was similar to the findings from experiment 1. Taken together, while there was support for retrieval ability as a predictor for modeling probability judgment sums in experiment 1, results from experiment 2 reveal inconclusive evidence on retrieval from semantic and episodic memory.

#### *Latent Variables Approach and Modeling Judgments*

Again, an impetus of experiment 2 is the use of a more intentional and refined latent variables approach by the use of structural equation modeling. One benefit modeling

judgments by use of factors or latent variables is that it allows for a cleaner test of the relation between the latent factors of interest and an outcome by accounting for measurement error. The results provided through Bayesian model comparison used composites to represent the unobserved memory abilities (WM, SM, and EM). Because composites do not account for task measurement error, the relation between composites and an outcome of interest may often be attenuated. The general philosophy taken in experiment 2 is to first, assess the memory measurement model. This is similar to the CFA conducted in experiment 1 where the 2 and 3-factor memory models were evaluated. Then, path analysis using a two-step structural equation modeling (SEM) procedure will be conducted on the appropriate measurement model to predict the different types of judgments. Altogether, this approach will allow us to provide converging evidence for the Bayesian model comparison results and assess how we are measuring what we hope to measure. More details on each of these methods will be outlined below.

#### Confirmatory Factor Analysis Method

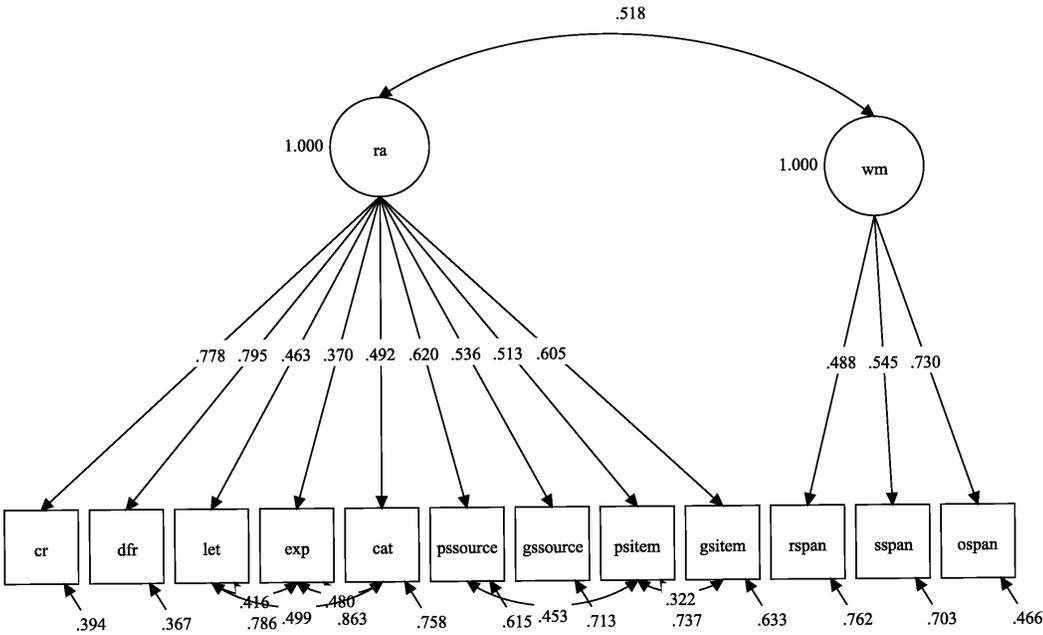
The goal of the CFA in experiment 2 is to ensure that the measures appropriately load onto their respective latent memory ability. Because we are interested in understanding how constraints in different aspects of memory uniquely contribute to judgments, it is important to also take into account the collinearity among the latent abilities. Further, while several models of memory have been proposed and researched, the theory behind why short and long-term memory tasks covary may differ depending on the memory model adopted. By using theory to guide different proposals of why certain tasks that tap into different aspects of memory correlate with

one another, one can gain insights into the more true relation between these memory abilities and how they impact certain outcomes of interest.

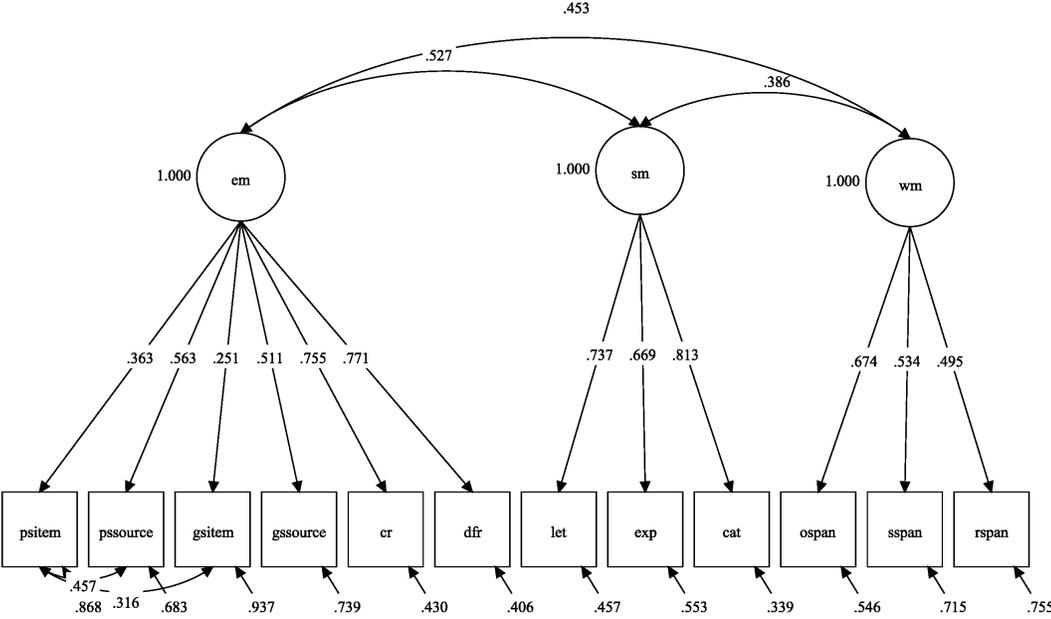
In the following analysis, four models of memory were assessed. The first is a 2-factor memory model. This is similar to how experiment 1 measured WM and RA, but in experiment 2, more tasks were added. The next model is a 3-factor memory model. This memory model parses RA into SM and EM. This was how composites were formed for Bayesian model comparison. Two additional memory models, a bi-factor and 2<sup>nd</sup> order model, were also evaluated. The bi-factor (residual) model is one in which there is a general retrieval ability that drives performance across all memory tasks, and then 3 other latent memory abilities that are unique to specific groupings of tasks (WM, SM, and EM). Finally, the 2<sup>nd</sup> order model is a model where SM, EM, and WM all depend on or is caused by a general latent retrieval ability that underlies the performance across all the memory tasks. In contrast to the bi-factor model, there is a direct causal path between RA and SM, EM, and WM.

The indicator variables used in the following CFA are individual performance or outcome scores on each task and additionally, were the variables used to compute the memory composites in the Bayesian model comparisons. CFA analysis followed a similar approach to experiment 1, in that error variances of indicator variables were allowed to co-vary and latent factors were also allowed to co-vary. Finally, again, the Satorra-Bentler Scaling method was used to adjust for non-normality (Satorra & Bentler, 2010).

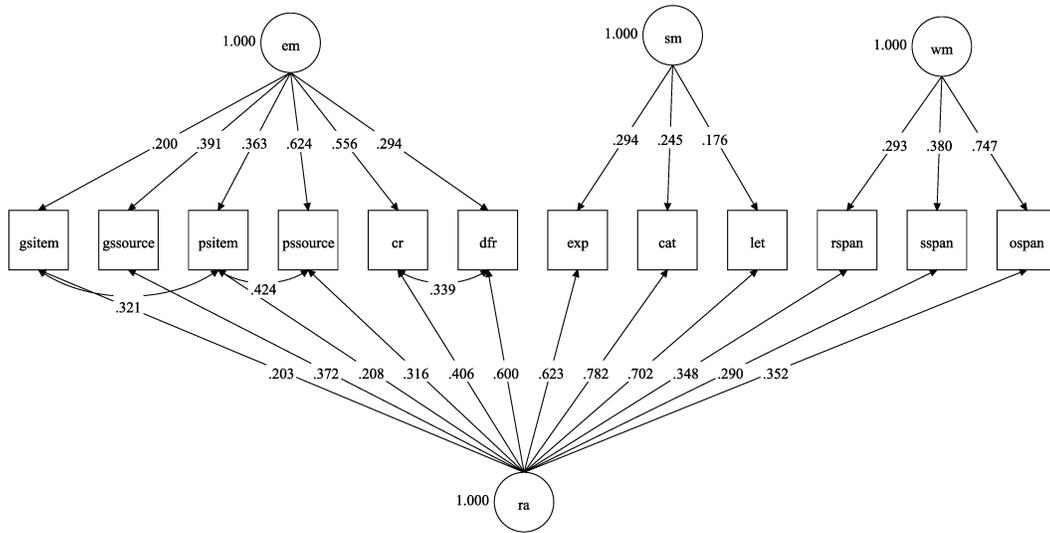
Confirmatory Factor Analysis Results



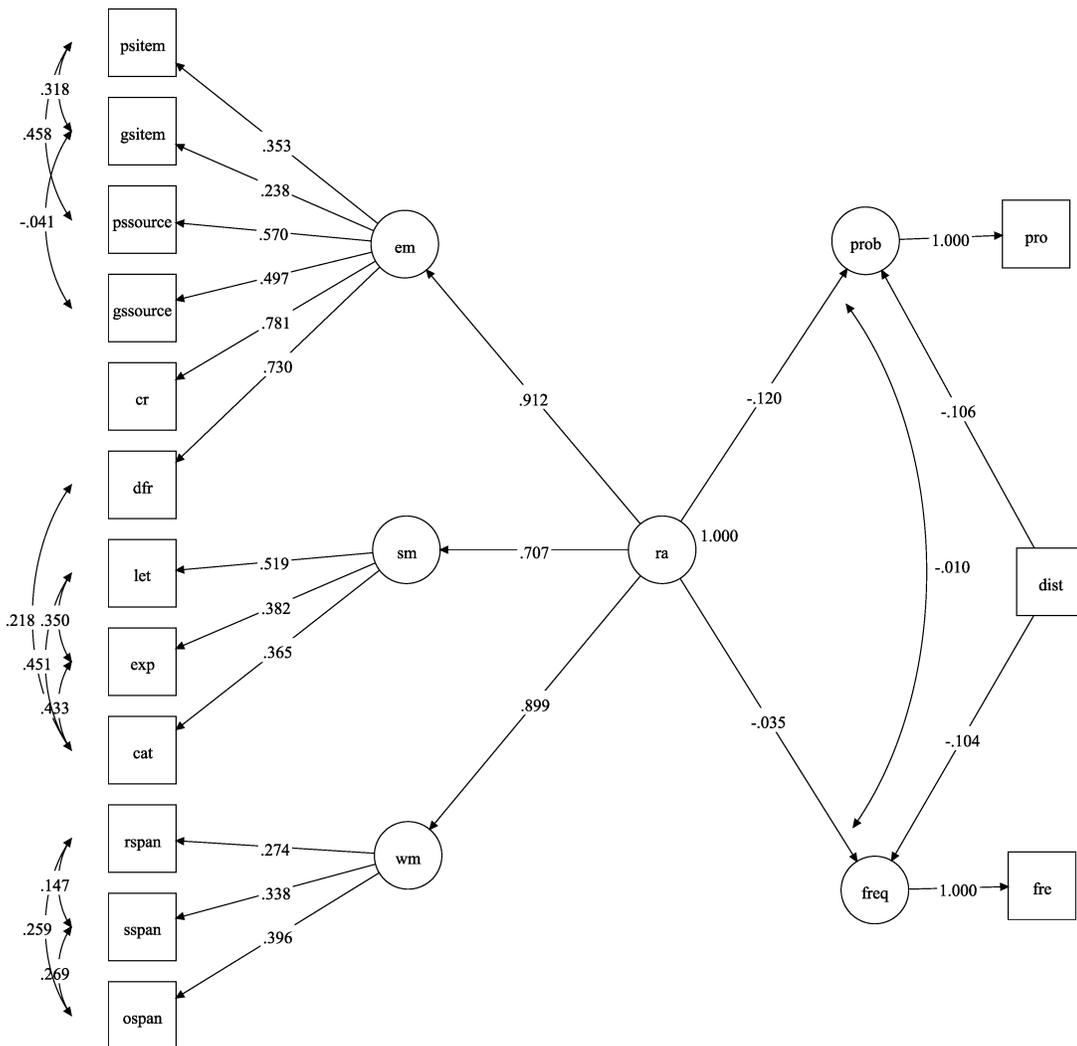
A. 2-Factor



B. 3-Factor



C. Bi-factor



D. 2<sup>nd</sup> Order

Figure 7. Standardized results from the CFA conducted on the 2-factor memory model (panel A), the 3-factor memory model (panel B), the bi-factor memory model (panel C), and the 2<sup>nd</sup> order memory model (panel D).

	<i>RMSEA with 95% CI</i>	<i>CFI</i>	<i>SRMR</i>	<i>BIC</i>
<i>Model A: 2 Factor</i>	<b>0.067</b> (0.04, 0.086)	0.921	0.146	11049.102
<i>Model B: 3 Factor</i>	<b>0.032</b> (0.00, 0.057)	0.982	0.043	11009.057
<i>Model C: Bi-factor</i>	<b>0.000</b> (0.00, 0.032)	1.000	0.027	11003.060
<i>Model D: 2<sup>nd</sup> Order</i>	<b>0.043</b> (0.015, 0.065)	0.967	0.068	11019.718

Table 3. Model fit indices of the four memory models from CFAs on experiment 2 data. Indices that meet the fit guidelines are bolded, with the exception of BIC.

Task reliabilities and construct quality indices are reported in the appendix. Figure 7 shows the standardized loadings of the 4 measurement models. All paths that are shown in this figure are significant. Table 3 shows model fit indices from the CFA. Again, using the measurement fit guideless proposed by Hu & Bentler (Hu & Bentler, 1995), the 3 factor, bi-factor<sup>3</sup>, and 2<sup>nd</sup> order memory models met the criteria for good data-model fit across the four memory models tested.

First, focusing between the 2 and 3-factor memory model, because the 2-factor model is nested within the 3-factor model, a chi-square difference test can be used to determine whether the addition of more parameters was worth the increase in overall data-model fit. Results revealed that the loss of degrees of freedom was worth the increase in fit,  $\chi^2(1) = 27.66901$ ,  $p = 7.44157e-08$ . Further, the 2<sup>nd</sup> order model is also nested within the bi-factor model (Yung, Thissen, & McLeod, 1999). Results

<sup>3</sup> Note that the fit indices for RMSEA and CFI on the bi-factor model should not be strictly interpreted as having perfect fit.

revealed that the loss of degrees of freedoms was worth the increase in fit,  $\chi^2(12) = 37.45156$ ,  $p = 0.00007$ .

Additionally, the BIC values were compared and BFs were computed to compare all four models. Results revealed strong evidence that the data are more likely to be observed under the 3-factor model over the 2-factor model (BF = 496205146). Between the bi-factor model the 2<sup>nd</sup> order model, the bi-factor model had strong support (BF = 4142.273). Finally, the bi-factor memory model was also preferred over the 3-factor model (BF = 20.05543). Taken together, of the four memory models, the bi-factor model appears to be the most appropriate measurement model.

#### Structural Equation Modeling Method

The purpose of using SEM is to evaluate the structural paths between each judgment type and memory ability. The central hypotheses and predictions still remain the same. Simply put, the prediction is that each judgment type will have a different relation between each memory ability. Further, with the assumption that frequency and probability judgments engage different cognitive processes, the expectation is that the relations between each judgment type and memory ability will differ as a function of the judgment type. Results from the CFA revealed that the 3-factor, bi-factor, and 2<sup>nd</sup> order memory models met measurement fit guidelines. Thus, path analysis will continue with these models. With these models, the interest will be to evaluate the direct causal paths between probability judgments and each memory ability and frequency judgments and each memory ability.

A 2-step SEM procedure will be used where the measurement model is first assessed, and then the structure of the latent model is evaluated. The purpose of the two-step SEM process is to isolate the data-model fit associated with the latent structure from that associated with the measurement portion of the model. The first step is one where all factors and/or standalone variables that are not intended to be indicator variables are allowed to co-vary. This is often referred to as the measurement phase. Note that the models tested during this measurement phase are different from the philosophy of the CFAs conducted on the measurement models. This is because those purpose of the CFAs was to assess only the latent structure of memory. Thus, these models did not include the outcome variables (judgment sums). The second step of the 2-step SEM procedure is one where a-priori paths are imposed onto the final measurement model that resulted from the measurement phase. This is often referred to as the structural phase. The data-model fit after the theoretical structure is imposed can then be compared to the data-model fit of the final measurement model.

#### Structural Equation Modeling Results

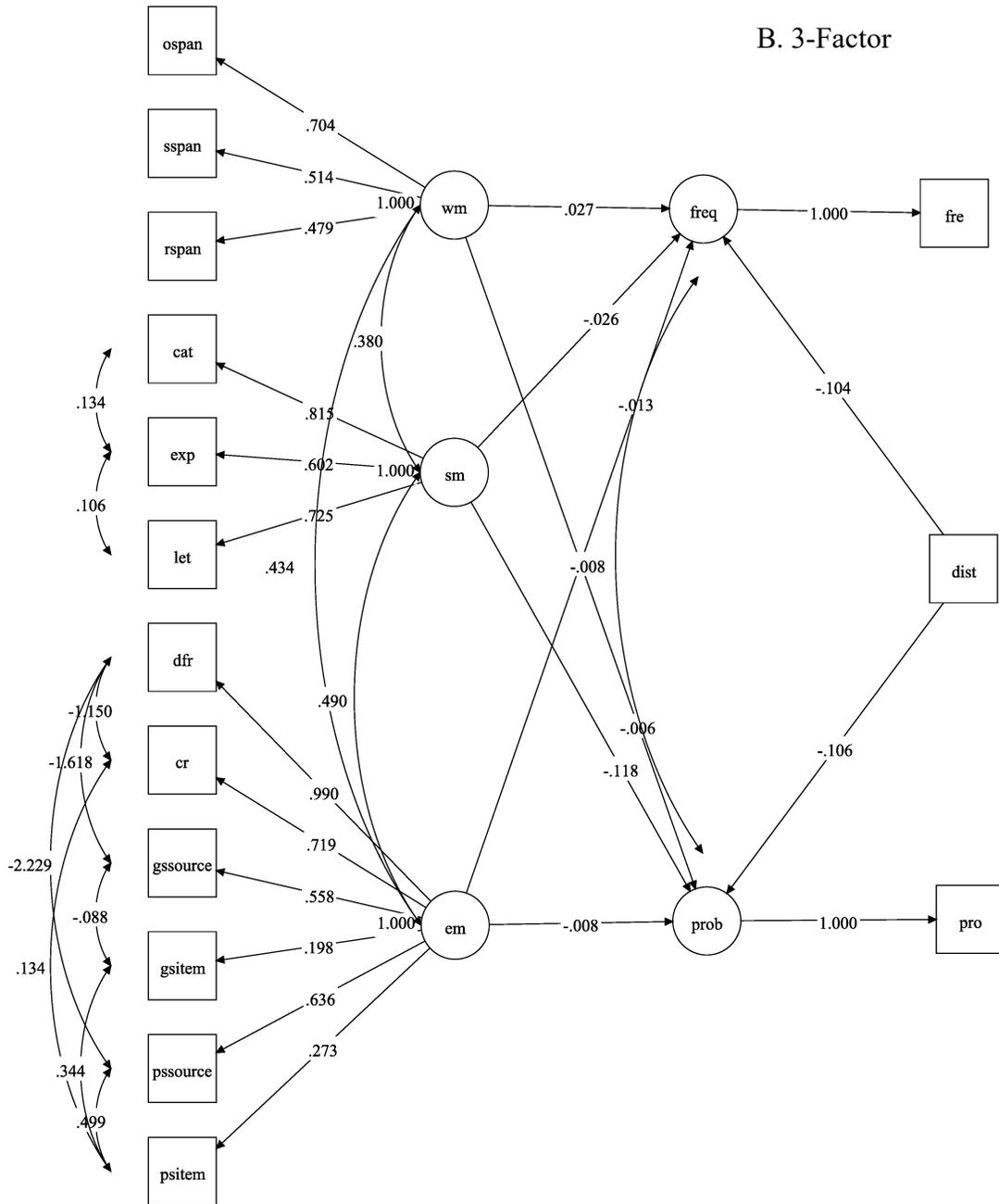
	<i>RMSEA with 95% CI</i>	<i>CFI</i>	<i>SRMR</i>	<i>BIC</i>
<i>Model B: 3 Factor</i>	0.037 (0.024, 0.049)	0.967	0.036	35150.699
<i>Model C: Bi-factor</i>	0.049 (0.037,0.060)	0.947	0.047	35189.747
<i>Model D: 2nd Order</i>	0.043 (0.031, 0.054)	0.951	0.050	35162.279

Table 4. SEM fit indices of the final structural models from experiment 2.

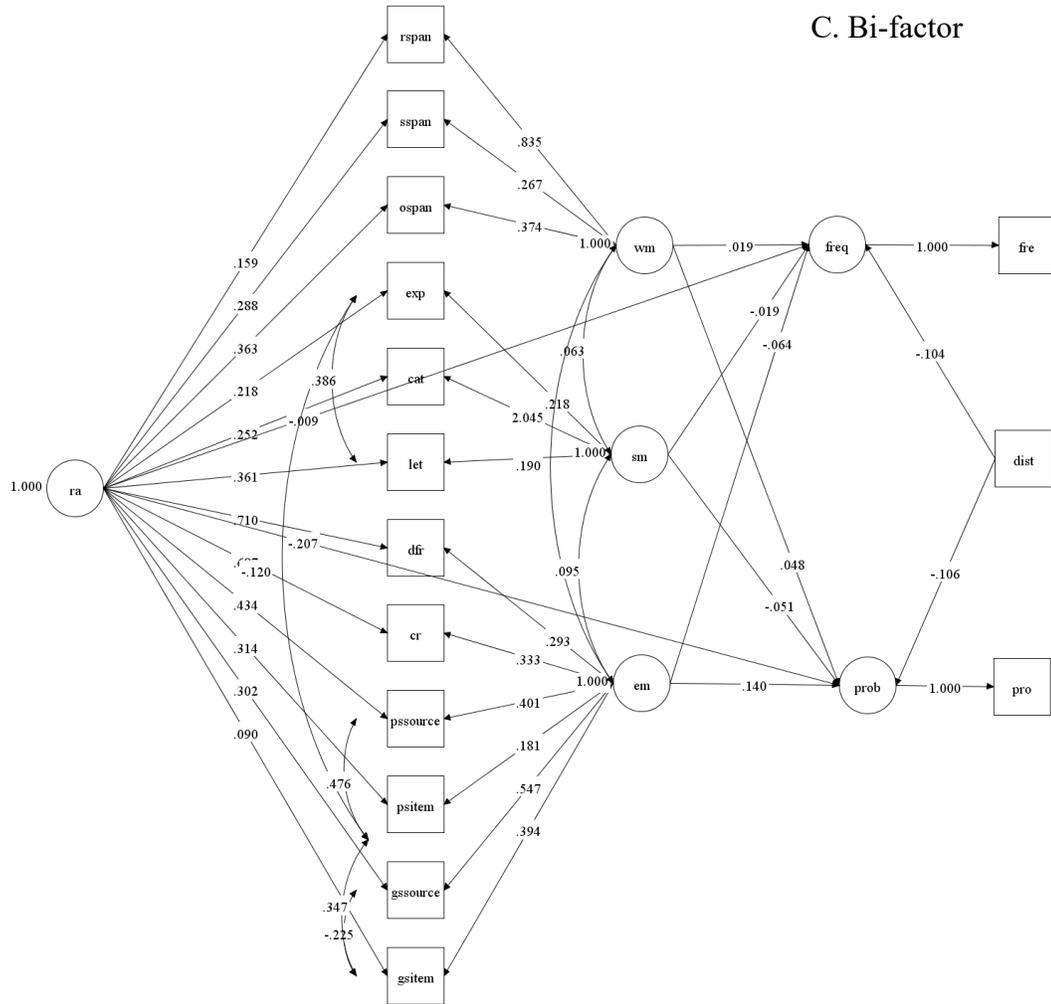
Table 4 provides the fit indices of the final measurement models resulting from the final structural phase. Because the paths between the measurement phase and

structural phase were the same, the fit indices remained the same. As can be seen, all three models that met satisfactory fit from the CFA also met satisfactory fit criteria from the final structural portion. The model with the lowest BIC was the 3-factor model. This also seemed to be the case when comparing the model fit indices across the three models. Again, the BIC values were compared and BFs were computed to compare all 3 models. Results revealed evidence that the data are more likely to be observed under the 3-factor model over the bi-factor model (BF = 301415419) and the 2<sup>nd</sup> order model (BF = 327.013). Between the bi-factor and 2<sup>nd</sup> order model, results showed evidence that data are more likely to be observed under the 2<sup>nd</sup> order model (BF = 921723).

B. 3-Factor



### C. Bi-factor



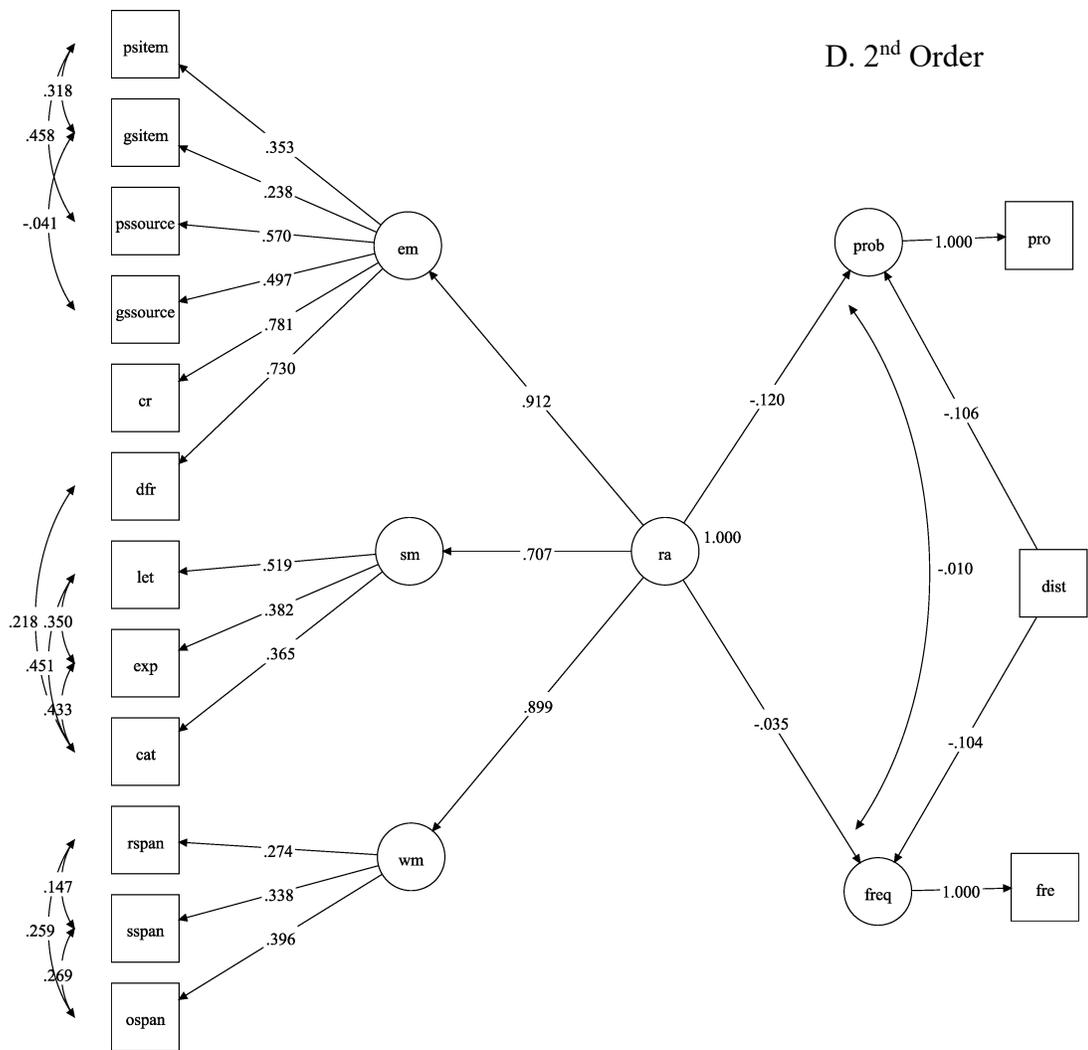


Figure 8. Structural path models with standardized results for the 3 factor (B), bi-factor (C), and 2<sup>nd</sup> order (D) models.

<i>Memory Model</i>	<i>Path</i>	<i>Estimate</i>	<i>P-Value</i>
3 Factor	Probability Judgments ON		
	SM	-0.026	0.635
	EM	-0.013	0.811
	WM	0.027	0.638
	Frequency Judgments ON		
	SM	-0.118	0.082
	EM	-0.008	0.882
	WM	0.022	0.947
Bi-factor	Probability Judgments ON		
	SM	-0.051	0.018
	EM	0.140	0.058
	WM	-0.207	0.001
	RA	0.048	0.457
	Frequency Judgments ON		
	SM	-0.019	0.305
	EM	-0.064	0.549
	WM	-0.009	0.746
	RA	0.019	0.895
2 <sup>nd</sup> Order	Probability Judgments ON		
	RA	-0.120	0.019
	Frequency Judgments ON		
	RA	-0.035	0.488

Table 5. Direct effect estimates for the 3 factor, bi-factor, and 2<sup>nd</sup> order models from experiment 2 data.

Figure 8 shows the results of path analysis conducted on the 3 models. Table 5 provides the direct effect estimates along with their associated p-value between two judgment types and latent memory abilities, or the paths of interest. For the 3 factor model, none of the paths between the latent memory abilities and judgments were significant. This provides converging evidence with the Bayesian model comparison results from experiment 2 data. When examining the direct effect estimates in the bi-factor model, the significant paths were ones between SM and probability judgments and RA and probability judgments. Finally, in the 2<sup>nd</sup> order model, the significant path is between RA and probability judgments.

### Experiment 2 Discussion

In experiment 2, Bayesian model comparison results revealed evidence for the main effect of distribution and judgment type. For both probability and frequency judgments, results only supported distribution to be included in modeling both types of judgments. Further, results were inconclusive for each memory ability in predictive both probability and frequency judgments. Further, CFA revealed that out of the four memory models assessed, 3 models met satisfactory fit, the 3-factor model, bi-factor model, and 2<sup>nd</sup> order model. Path analysis conducted on these models using a 2-step SEM procedure revealed interesting results. On the 3 factor model, none of the paths between the latent memory abilities and each judgment type were significant. For the bi-factor model, SM and RA were predictive of probability judgments. Finally, for the 2<sup>nd</sup> order model, general RA was significant in predicting probability judgments. Taken together, this suggests that how memory is modeled and measured may influence which memory abilities appear to meaningful in predicting judgment sums.

## Chapter 6: Combined Data

Results from experiment 2 reveal that how memory is modeled and measured may influence which latent abilities come out as important to modeling judgments. The purpose of combined data analysis is to gain additional power. Additionally, the central questions of interest were re-analyzed using both the 2 and 3-factor memory model on predicting different judgment types to examine the robustness of

experiment 1 and 2 Bayesian model comparison findings and will also be the source of the overall project's general conclusions.

As mentioned previously, model comparison or selection typically tells one which model best fits the data *relative* to the other models of which it is compared. Further, in the philosophy that a model's utility lies in its predictive ability, cross-validation on an overall model, and each judgment type will be performed. The thinking is that if a model has high predictive quality, then the new data predicted from the model fitted with the observed data and thus, both predicted and observed data for within and out of sample should be similar to one another. This can be conceptualized as the model's ability to predict unobserved data. Finally, the magnitude of BFs from hypotheses testing for each predictor should not be conflated with an independent variables effect estimate. Since cross-validation will be performed, parameter estimates for each predictor of interest will be estimated and reported.

### Analysis

#### Coding of Variables and Constructs

The variables that were the same in both experiment 1 and 2 were included in the combined analysis. The data set used in the analysis reported below contained raw scores from each experiment. Once the data were combined across both experiments, z-scores and composites were re-calculated using the means and standard deviations from the combined sample. The same method of forming composites used in experiment 1 and experiment 2 were used for the combined dataset. This includes the

handling of missing data points, outliers, and the re-calculation of z-scores and composites on data that did not contain those who fell below threshold performance on the complex span tasks.

Further, similar to experiment 2, a 3-factor memory model was used to examine the impact of individual differences in memory on judgment sums using Bayesian model comparison. For the combined dataset, the memory abilities are represented by composites constrained by measures that were included in both experiments. Thus, WM was measured by operation and symmetry span, SM was measured by category and experience fluency, and EM was measured by DFR. Finally, all model comparisons controlled for the random effect of participant and experiment.

## Results

### Final Analysis Sample and Demographics

All participants who were included in analyses from experiment 1 and experiment 2 were included in the combined analysis. Thus, the final sample size of the combined dataset included three-hundred and seventy-five participants (262 females, mean age =  $19.10 \pm 2.03$ ). The total number of participants for the dataset containing only those who met performance criteria on working memory measures was 311. The number of participants in the dataset with potential outliers removed was 335, and the number of participants in the dataset with potential outliers removed and who also met performance criteria on working memory measures was 295.

## Descriptives

A correlation matrix of the judgments, composite memory abilities, and task variables are located in the appendix.

## Question 1

*Is there evidence of the comparison process for probability judgments and lack of one for frequency judgments?*

Again, the prediction for this question is that the overall magnitude of probability judgment sums for evenly distributed items would be greater when compared to the magnitude of unevenly distributed. This is not expected for frequency judgment sums as a function of distribution.

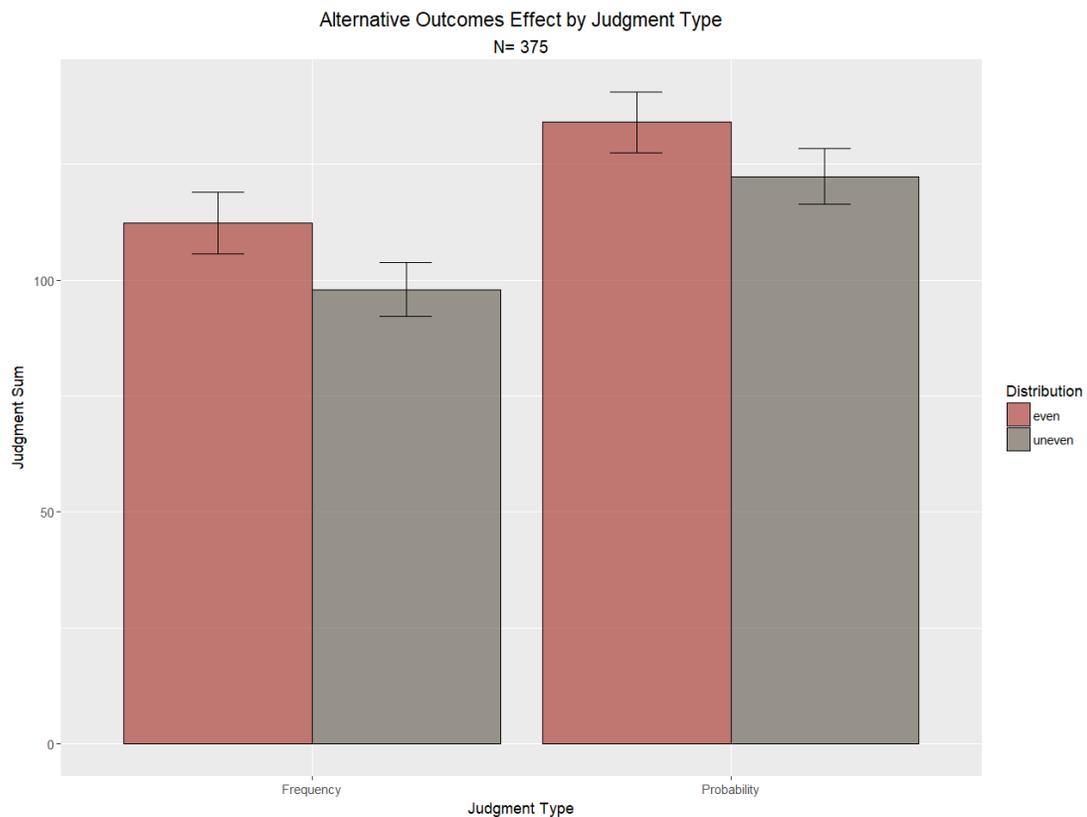


Figure 9. The effect of distribution on judgment sums by judgment type. The error bars represent 95% confidence intervals.

Figure 9 shows judgment sums as a function of judgment type and distribution and again, the figure shows that the overall judgment sums for items that were evenly distributed are greater when compared to those from the uneven distribution. The difference between findings from experiment 1 and 2 was that experiment 1 had inconclusive evidence of distribution as a main effect and for probability judgments, while experiment 2 found support for distribution in both models. Analysis on the combined dataset revealed strong support for judgment type ( $r = 0.149, 9598575966 \pm 2.06\%$ ) and distribution ( $r = -0.084, 350.252 \pm 2.15\%$ ). Additionally, the test of the interaction model with the two-main effects model revealed that the data are over 8 times more likely to be observed under the 2 main-effect model over the interaction model ( $BF = 0.127 \pm 2.24\%$ ). Finally, there was decisive support for the effect of distribution for frequency ( $BF = 216253820269 \pm 7.74\%$ ) and probability ( $BF = 945919.6 \pm 7.32\%$ ) judgments sums.

## Question 2

*What is the relation between each judgment type and memory ability?*

The findings from experiment 1 suggest that those who perform better on long-term memory retrieval measures had less subadditive judgments when compared to those who did not perform as well. Further, exploratory analysis suggests that semantic fluency measures may be the driving force behind this finding. For the current experiment, retrieval ability was parsed into semantic memory (SM) and episodic memory (EM). The prediction is that RA would still be important to modeling probability judgment, but that SM would have the strongest support among the

retrieval abilities. Figure 10 shows the relation between probability judgment sums and each memory ability.

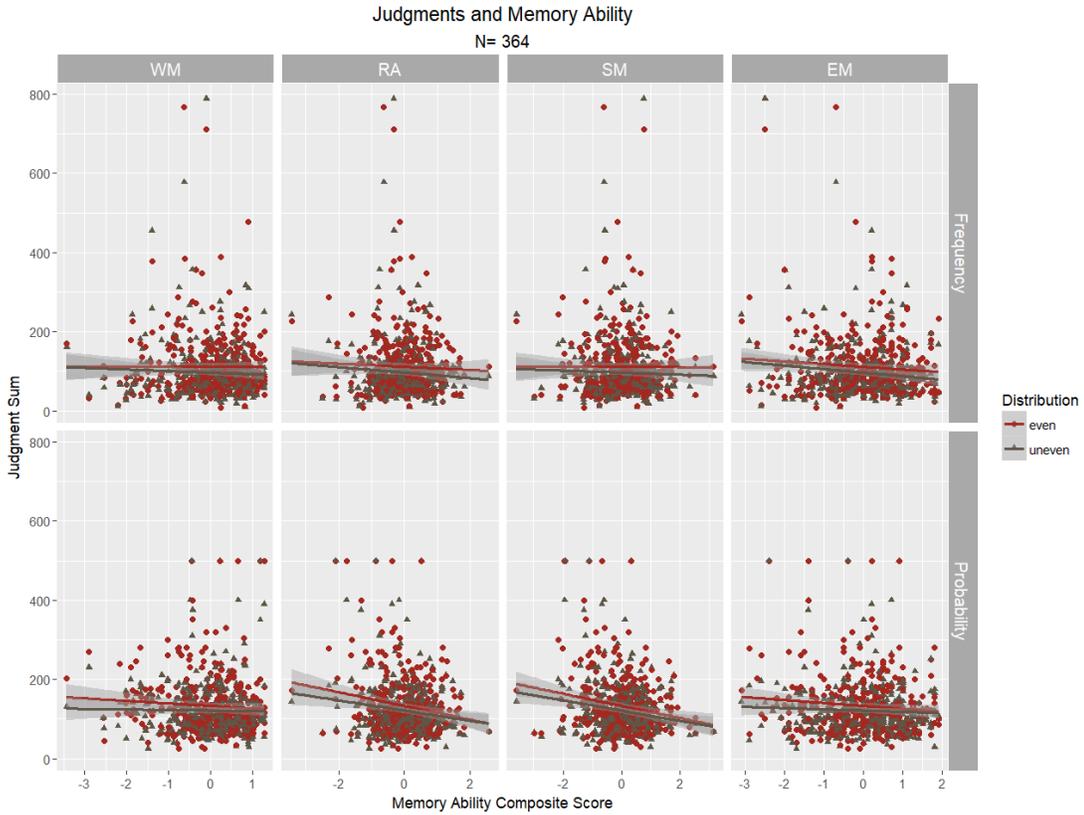


Figure 10. The relation between memory ability and judgment sums as a function of judgment type. The grey area represents the 95% confidence interval region.

	Distribution	Working Memory	Retrieval Ability
<i>Probability</i>			
Experiment 1	0.37±4.42%	0.70 ±4.79%	27.78 ±5.85
Experiment 2	18747146 ±6.59%	0.52 ±4.94%	1.032±6.5%
Combined	531340.8 ±7.94%	0.32 ±7.92%	18.65±8.8%
<i>Frequency</i>			
Experiment 1	10093.86 ±6.12%	0.68 ±5.14%	1.19 ±6.22%
Experiment 2	11751359 ±7.24%	0.51 ±7.07%	0.43 ±6.87%
Combined	628897393188 ±9.88%	0.54 ±9.29%	0.69 ±9.21%

Table 6. BFs for Memory abilities predicting each judgment sum type across each set of analyses. BFs for retrieval ability was assessed using a 2-factor memory model.

	<i>Distribution</i>	<i>Working Memory</i>	<i>Semantic Memory</i>	<i>Episodic Memory</i>
<i>Probability</i>				
<i>Experiment 1</i>	0.42 ±3.86%	0.53 ±6.22%	38.74 ±6.06%	0.48 ±4.22%
<i>Experiment 2</i>	17972097 ±8.41%	0.63 ±8.07%	0.94 ±8.09%	0.58 ±8.07%
<i>Combined</i>	863037.2 ±8.35%	0.46 ±8.08%	21.67 ±8.97%	0.49 ±8.1%
<i>Frequency</i>				
<i>Experiment 1</i>	9325.93 ±20.15%	0.94 ±21.32%	0.56 ±21.75%	18.80 ±20.4%
<i>Experiment 2</i>	5508386 ±9.85%	0.57±8.95%	0.53 ±8.96%	0.56 ±8.93%
<i>Combined</i>	640948315656 ±9.85%	0.50 ±9.21%	0.56 ±9.28%	2.44 ±9.56%

Table 7. *BFs for Memory abilities predicting each judgment sum type across each set of analyses. BFs for retrieval ability was assessed using a 3-factor memory model.*

The scatterplot reveals negative correlations between each memory ability and judgment sum, regardless of judgment type. When focusing on a 2-factor memory model, the strongest correlation to both types of judgments was RA (frequency judgments:  $r = -0.057$ ; probability judgments:  $r = -0.165$ ). For WM, the correlation was  $r = -0.046$  for probability judgments and  $r = -0.024$  for frequency judgments. For the 3 factor model, probability judgments have the strongest correlation was with semantic memory ( $r = -0.177$ ), followed by episodic memory ( $r = -0.073$ ), and lastly, working memory ( $r = -0.045$ ). For frequency judgments, the strongest correlation was with retrieval from episodic memory ( $r = -0.102$ ), followed by working memory ( $r = -0.022$ ), and then finally, semantic memory ( $r = -0.017$ ).

Table 6 and 7 show the BFs of each memory ability on modeling each judgment type by the data source. Table 5 uses a 2-factor memory model and table 6 uses a 3-factor memory model. It is worth noting that the memory composites used in these models are constrained to the 5 tasks that were included across both experiments (category fluency, experience fluency, delayed-free recall, operation

span, and symmetry span). When considering the 2 factor model on the combined data, there is strong support for the inclusion of distribution when modeling judgment sums. This was also the case for RA in modeling probability judgments.

Considering the 3-factor memory model, again, there is strong support for the inclusion of distribution in modeling judgment sums. Further, individual differences in semantic memory had support for being included in the model predicting probability judgment sums. For frequency judgments, results were generally inconclusive for each memory ability across each set of analysis.

### Assessing Judgment Model Utility

*Overall:*

*Judgment ~ Judgment Type \* Distribution + WM + SM + EM*

*By Judgment Type:*

*Probability Judgment ~ Distribution + WM + SM + EM*

*Frequency Judgment ~ Distribution + WM + SM + EM*

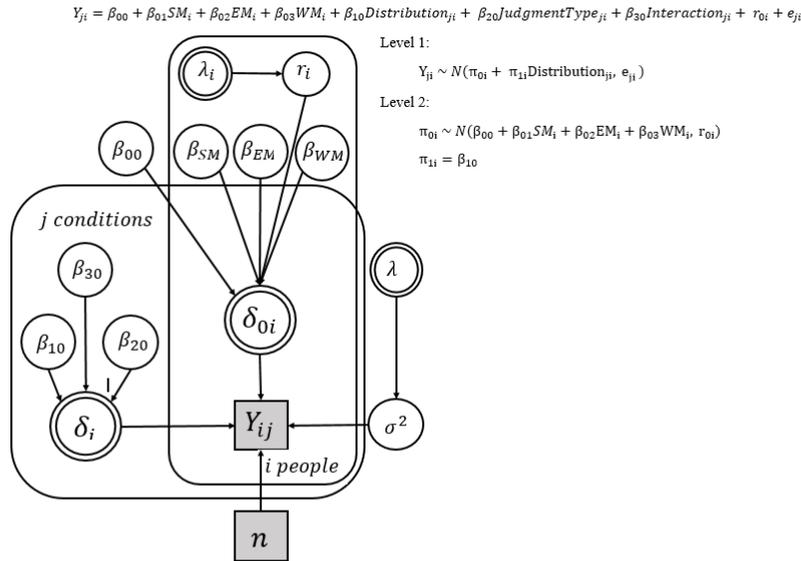
Bayesian parameter estimation and cross-validation will be assessed for each of the models above.

### Cross-Validation and Parameter Estimation Methods

*Modeling Judgments.* Theoretically, the outcome variable, judgment sums, is a function of each individual's starting point (intercept), the overall main effect of the condition, and overall error. The design of the study is hierarchical where level 1 examines the effect of the condition (evenly distributed items versus unevenly

distributed items) and level 2 is at the level of participant, where intercepts are allowed to vary as a function of an individual's latent memory ability.

### A. Overall



### B. By Judgment Type

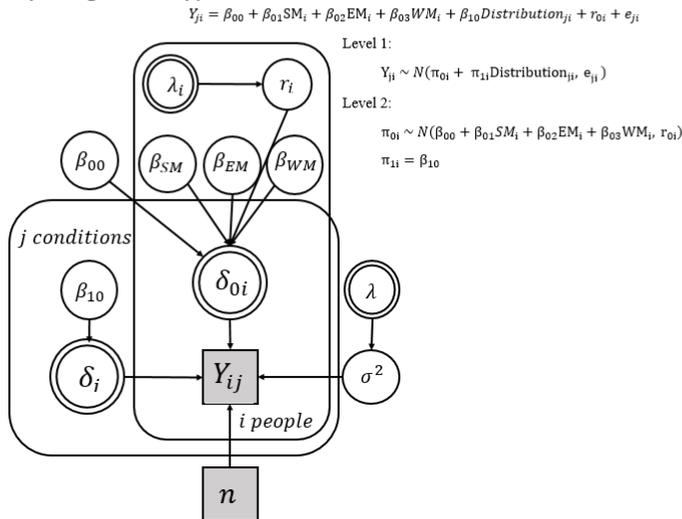


Figure 11. Graphical representations of the cognitive models for overall judgments (A) and for each judgment type (B).

The graphical models in figure 11, along with the overall and level 1 and 2 equations shows what each parameter is a function of. The top panel shows the cognitive model for overall judgments and the bottom panel shows the cognitive model used in

predicting each judgment type (probability and frequency). Parameters in double circles are deterministic. For example, the delta of the condition is determined by the effect of condition, which is the difference between the two judgment sums means of each distribution. The delta of the intercept is determined by the overall intercept, the beta estimate of retrieval ability, the beta estimate of working memory, and the error of each participant. For the overall model, there will be two additional level 1 terms, judgment type ( $B_{20}$ ) and the interaction between judgment type and distribution ( $B_{30}$ ).

*Specification of Priors and Likelihoods.* For all parameters that were estimated, weakly informative priors were used. All betas were a function of a wide uniform distribution. The level 2 error (the effect of participant), the assumption is that each individuals effect came from the same normal distribution with a mean of 0 with some unknown precision. The precision at level 1 and level 2 are independent of each other, but at each level, are drawn from the same uniform distribution truncated at 0. The specific values of each distribution are presented below.

<i>Parameter</i>	<i>Prior Specification</i>
<i>Fixed effects/ Betas parameters</i>	<i>betas and interactions <math>\sim U(-1000,1000)</math></i>
<i>Random Effect parameters</i>	<i><math>r_i \sim N(0, \lambda_i^2)</math>, where <math>\lambda_i^2 = 1/\sigma_i^2</math></i>
<i>Variance parameters</i>	<i><math>\sigma_j^2, \sigma_i^2 \sim U(0, 1e + 7)</math></i>

*Table 8. Table of priors placed on model parameters to be estimated.*

*Sampling Procedure.* Sampling was performed within R version 3.5.0 and the models were written in JAGs version 4.3.0. The number of chains was set to 5 at 50,000 iterations each. The number of adaptations and burn-ins was set to 10,000. All initial convergence assessment procedures include the visual assessments of trace, density, running means, autocorrelation, and potential scale reduction factor plots. Initial assessment of the sampling procedure with the aforementioned sampling

parameters showed high autocorrelation so another model was initiated with thinning of 5 at 50,000 iterations for a total of 10,000 chains. However, the autocorrelation and all estimates were still similar. Because Monte Carlo Markov Chain (MCMC) was used to get estimates of effects and perform cross-validation, typical convergence and model diagnostics were assessed using standard practices (ex: trace plots, running means plots, posterior scale reduction factor, etc.). These evaluations are provided in the appendix. Additional plots or supplements can be provided upon request.

*Cross-Validation Procedure.* Each model was fitted with 2/3<sup>rd</sup> of the sample.

Selection of the sample was randomized for each model. The parameter estimates were obtained and were used to predict within sample (the sample that the estimates were fitted to) and out of sample (the remaining 1/3<sup>rd</sup> of the sample). The parameter estimates used to fit the two-thirds of the sample will be reported.

## Cross-Validation and Parameter Estimation Results

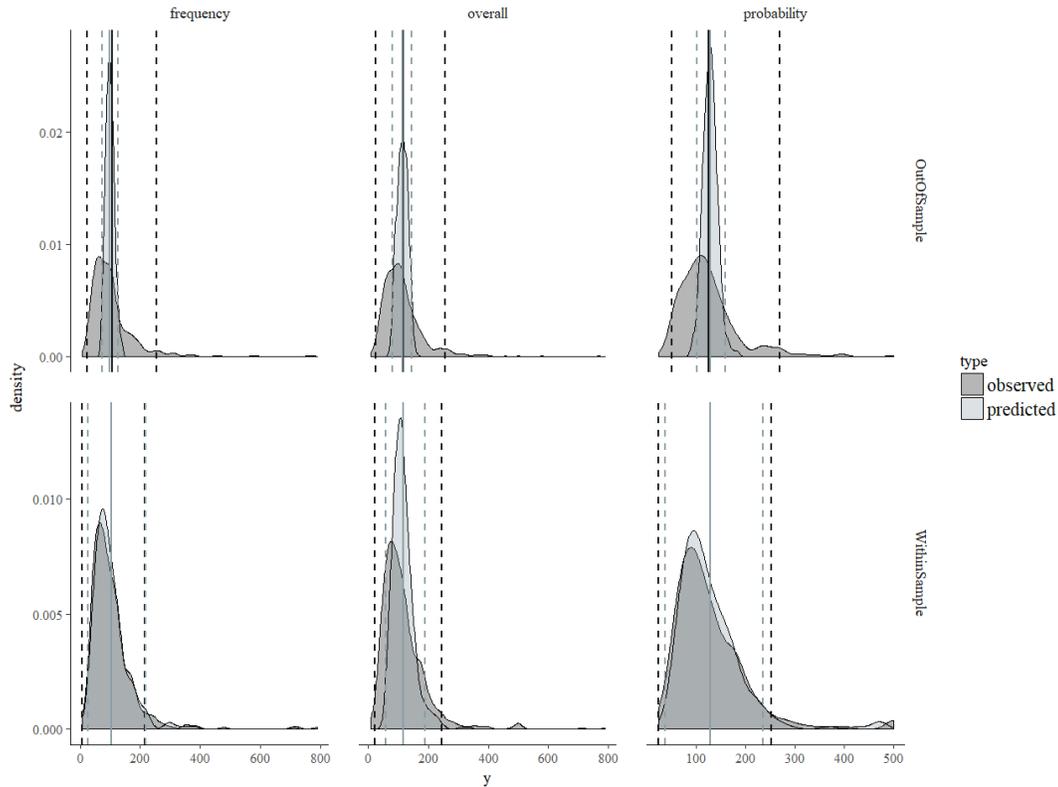


Figure 12. The distribution of the observed and predicted values from the overall, probability, and frequency judgment models by prediction type. The dashed lines represent the 95% HPD interval.

Figure 12 is the visual result of each model's within and out of sample prediction.

For each model and prediction type, the distributions of the observed and predicted Ys are overlaid. Each distribution is also plotted with its 95% high posterior density (HPD) interval. This allows one to assess how similar the observed and predicted distributions of the judgment sums are. Ideally, the distributions should be similar in shape. Further, the predicted values should lay within the 95% HPD interval of the observed values. Visually, across all three models, the observed and predicted distributions for both within and out of sample predictions seem very similar. Additionally, 95% of the predicted data seem to be contained within much of the observed data.

<i>Model</i>	<i>Prediction</i>	$\bar{Y}$	$\hat{Y}$	<i>MSE</i>
<i>Overall</i>	<i>Within</i>	128.55	128.54	151783372
	<i>Out</i>	125.02	128.23	638754187
<i>Probability</i>	<i>Within</i>	115.62	115.62	924694
	<i>Out</i>	115.72	113.38	139345576
<i>Frequency</i>	<i>Within</i>	102.69	102.71	151783372
	<i>Out</i>	106.42	98.74	1162715653

Table 9. Descriptives of the distribution of outcome values for within and out of sample prediction on each model.

Table 9 shows the descriptives of the outcome variable for within and out of sample predictions for all 3 models. The discrepancy between the mean squared errors (MSEs) for within sample can be compared to the MSEs for out of sample prediction, which is often referred to as the mean squared predicted errors (MSPEs). The ratio between these numbers suggests that the overall model was the least discrepant, followed by the probability model. The frequency model appears to be the model most subject to possible overfitting of the data.

<i>Model</i>	$R^2$	<i>Judgment Type</i>	<i>Distribution</i>	<i>Interaction</i>
<i>Overall</i>	0.33 (0.295, 0.365)	26.8 (15.8, 37.8)	16.4 (5.2, 27.3)	-1.8 (-17.5, 13.4)
<i>Probability</i>	0.86 (0.836, 0.873)	-	14.56 (9.5, 15.6)	
<i>Frequency</i>	0.89 (0.882, 0.909)	-	16.33 (12.0, 20.5)	

<i>Model</i>	<i>Working Memory</i>	<i>Semantic Memory</i>	<i>Episodic Memory</i>
<i>Overall</i>	3.4 (-5.2, 12.0)	-5.6 (-13.4, 2.0)	-7.1 (-14.9, 0.6)
<i>Probability</i>	2.3 (-9.9, 14.2)	-14.93 (-25.9, -4.2)	-1.7 (-12.7, 8.9)
<i>Frequency</i>	4.6 (-7.5, 16.8)	3.8 (-7.6, 16.8)	-12.6 (-23.6, -1.7)

Table 10. Median effect size estimates with 95% HPD interval from the posterior distribution using a 3-factor memory model.

Finally, the parameter estimates from the fitted samples are shown in table 10. Each parameter estimated includes the median value with its 95% HPD interval. Notice that for the overall model, only judgment type and distribution do not contain 0 within the HPD interval. For probability judgments, distribution and semantic memory were the only parameters that did not include 0 within the HPD interval. Finally, for frequency judgments, distribution and episodic memory were the parameters that did not contain 0 within the HPD interval. This somewhat converges with prior Bayesian model comparison findings. Taken together, this suggests that the variance across all judgments are a function of the type of judgment and the distribution of which it came from. For probability judgments, the data are more likely to be observed under models with SM included as a predictor. For frequency judgments, though Bayesian model comparison results were inconclusive to weak, parameter estimation seems to indicate that EM may play a role in predicting frequency sums.

### *Combined Analysis Discussion*

Results from Bayesian model comparison addressing the central core questions reveal several things. With judgments predicted using a 2-factor memory model, results show support for RA in predicting probability judgment sums. When judgments are predicted using a 3-factor memory model, the data for probability judgments are more likely to be observed when SM is included as an independent variable. Across all judgment sums, both distribution and judgment type had strong evidence for predicting judgment sums. Further, parameter estimates obtained from the cross-validation procedure on all three judgment models provided converging and insightful

information. Cross-validation of each main effects model (overall, probability, and frequency judgment sums) seem to show good predictive quality for within sample prediction and decent predictive quality for out of sample prediction.

## Chapter 7: End Remarks

### What Have We Learned?

In experiment 1, results revealed that retrieval from long-term memory may play an important role in predicting probability judgments sums. Exploratory analyses suggested that general retrieval ability and its relation to judgment sums may vary as a function across judgment type and also within retrieval ability.

Confirmatory factor analysis showed that a 3-factor memory model may be a more appropriate way to represent latent memory abilities.

In experiment 2, general retrieval ability was parsed into semantic and episodic memory. The replicability of the findings from experiment 1 showed that again, judgment type was important in predicting judgment sums. Additionally, the distribution of the alternatives choices also plays a role in both probability and frequency judgment. More measures were added in order to assess how memory may be best measured and to also take a latent variables modeling approach in predicting judgment sums. Confirmatory factor analysis revealed that a 3-factor memory model is appropriate. Structural equation modeling also suggested that the relation between latent memory abilities and judgments may be defined by which type of memory model one adopts. When memory was measured using a 2<sup>nd</sup> order model, retrieval ability predicted probability judgment sums. When memory was measured using a

bi-factor model, semantic and retrieval ability predicted probability judgment sums. Specifically, the better the individual performed on these memory tasks, the overall magnitude of the judgment sums were more accurate.

Finally, combining data across both experiments provided converging evidence. Two-factor memory models suggest that retrieval ability contributes to predicting probability judgment sums, and three-factor memory models revealed evidence for semantic memory predicting probability judgment sums. Model utility was evaluated through cross-validation. Bayesian estimation was used to obtain parameter estimates of predictors. These estimates of the effects of independent variables across the overall, probability, and frequency judgment models showed good within sample prediction and decent out of sample predictions.

#### *Implications of the Distribution of Items on Judgments*

One of the surprising and consistent findings in this project was the support of the main effect of distribution on both judgment types. The effect of the strength of evidence was expected to influence probability judgments but not frequency judgments. To gain some insights as to why this effect was found for frequency judgments, first we turn to prior research and then theory.

The hypothesis that there should be no support for the effect of the distribution of items on judgment sums for frequency judgments stemmed partly from prior work on differences between probability and frequency judgments (Sprenger & Dougherty, 2006). Here, Sprenger and Dougherty found that the alternative outcomes effect was significant for probability judgments but not for frequency judgments. They, therefore, concluded that probability and not frequency judgments entail a

comparison process. Interestingly, using the same paradigm and distributions as that study, the data from this current project did show the same pattern, as seen in table

11.

Mean judgments as a function of judgment type and distribution type				
	Experiment 1 & 2 Combined		Sprengr & Dougherty (2006)	
Distribution Type	Mean frequency estimate	Mean probability estimate	frequency estimate	Mean probability estimate
Frequency of Presentation				
<i>30-15-15-15-15 even distribution</i>				
30	28.50 (1.16)	34.50 (1.05)	25.63 (1.37)	44.73 (2.64)
15	20.63 (0.99)	25.21 (0.95)	17.59 (0.95)	26.85 (2.04)
15	20.68 (0.88)	24.55 (0.96)	17.26 (1.02)	26.21 (1.78)
15	21.43 (0.88)	25.85 (0.98)	16.51 (0.88)	24.99 (1.76)
15	21.07 (0.89)	23.94 (0.92)	16.81 (0.88)	23.16 (1.66)
<b>Sum</b>	<b>112.31 (4.21)</b>	<b>134.05 (3.98)</b>	<b>93.80 (4.11)</b>	<b>145.94 (7.56)</b>
<i>30-45-5-5-5 uneven distribution</i>				
30	29.26 (1.20)	35.78 (1.05)	26.15 (1.47)	38.96 (1.98)
45	34.35 (1.39)	44.23 (1.13)	36.15 (1.62)	55.96 (2.44)
5	11.11 (0.64)	14.24 (0.82)	7.32 (0.58)	10.17 (0.90)
5	12.11 (0.73)	14.42 (0.84)	7.40 (0.70)	11.79 (1.19)
5	11.16 (0.77)	13.72 (0.75)	7.13 (0.44)	12.23 (1.25)
<b>Sum</b>	<b>97.98 (3.93)</b>	<b>122.39 (3.61)</b>	<b>84.15 (3.48)</b>	<b>129.11 (5.45)</b>

Judgments are percentages. Standard errors are presented in parenthesis after the mean judgments.

Table 11. Mean judgments from data combined from experiment 1 and 2 compared to mean judgments from Sprengr and Dougherty (2006).

The alternative outcomes paradigm was used as a design choice because the assumption is that probability judgments require one to compare the strength of evidence of the focal item to an alternative item during the judgment process. In contrast, prior work on frequency judgments mainly point to factors such as representativeness, availability, and familiarity as the driving force behind their estimates. Why then, was there support for the distribution of the alternative items for frequency judgments? To address this, we turn to support theory (Tversky & Koehler, 1994), expressed by the following equation:

$$P(A, B) = \frac{s(A)}{s(A) + s(B)}$$

Simply put, support theory states that when making a likelihood judgment on A, that one will compare the strength of evidence of A over the strength of evidence of A plus the strength of evidence of the alternatives, which in this case is just B. As noted before, we assume a comparison process for judgments of likelihood as it inherent within the estimation process.

In the context of this experiment, we used a format that has been referred to as a natural sampling format (see Gigerenzer & Hoffrage, 1995). Thus, one can view that we examined frequency judgments by asking participants to estimate  $s(A)$  and assessed probability judgments by asking participants to estimate  $s(A)/s(A)+s(B)$ . The strength of evidence,  $s$ , can be seen as the base-rate or frequency of presentation of a particular item.

Prior research suggests that the biases that are seen in probability judgments of the function of the number of alternatives one considers and that the number of alternatives one can consider is constrained by working memory capacity. However, let's append to this the idea that judgments of likelihood are also a function of it's the strength of evidence. That is, judgments of likelihood are a process of two algorithms, the estimation of the strength of evidence of the to-be-judged item,  $s(A)$ , and the number of alternative items one can consider  $s(B)$  where  $B$  contains the alternatives  $s(b_1) + s(b_2) + s(b_3) + s(b_4)$ . It could be that the frequency estimates,  $s(A)$ , are driven by some sort encoding parameter which may result in differences in the sums between the two distributions. Prior research has suggested several different accounts that are hard to disentangle. For example, primary bias predicts an

overestimation of small stimulus frequencies and underestimation of large stimulus values (Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978). Other work suggest that people take into account, their uncertainties and regress their estimates towards the mean (Hertwig, Pachur, & Kurzenhäuser, 2005). Our data are unable to distinguish between these two patterns. However, it is interesting to note that in almost all cases shown in table 11, the order of magnitude for each item level estimate is higher in probability estimates when compared to its respective frequency estimates. Thus, this may suggest that future work should consider that biases in frequency judgments may be driven by an algorithm that deals with the estimation of the strength of the item. For probability judgments, two cognitive processes may be involved by adding to the frequency estimate of the strength, which involves the number of the alternative items one can consider.

#### *Implications of Memory on Judgments Types*

Finally, we now discuss the implications of these findings in the context of memory and how that accounts for some of the phenomena. The relation of interest is the support of retrieval ability being predictive of probability judgments. Specifically, semantic fluency. Why may fluency matter to probability judgments? Again, let's consider the design of this project, as well as prior research that has been conducted in a similar fashion. Participants are asked to judge the likelihood of a certain item within its class. This class, or superordinate category, in this design contained exemplars that were highly semantically related. Thus, the ability to retrieve additional items could be facilitated by the semantic relatedness of the exemplars within a particular class. However, note that for frequency judgments, there was

inconclusive evidence on semantic fluency. This may suggest that different cognitive processes are elicited by judgment type. Future studies can examine whether varying the semantic relatedness of a category of items that one will later judge would lead to more normative judgments. That is, it could be that if there is low semantic relatedness between exemplar items within a category that the retrieval of the alternatives may be more difficult, leading to more subadditive judgments. On the other hand, if the category contains highly related items, a judge may be able to retrieve a more complete set of alternatives, allowing for a more normative problem space and thus, more accurate judgment sums.

### Conclusions

Taken together, findings suggest that the type of judgment and the distribution of the alternative choices of the to-be-judged items can influence one's judgment estimate. Further, for probability judgments, individual differences in retrieval ability, specifically, semantic memory, seems to play a role in predicting judgment sums. Prior work has shown that frequency judgments are also subject to biases (Hertwig et al., 2005; Lichtenstein et al., 1978). Our results from Bayesian model comparison were generally inconclusive in regards to each memory ability. Future work should examine the underlying cognitive mechanism that may drive these biases.

In so far that researchers seek to understand these phenomena in order to provide prescriptive models, the re-framing of phenomena, even if only descriptive, can help move the field forward. Just as the judges in our experiments are assumed to want to break down complex problems by use of simple heuristics, researchers are also trying to break down the complex inner workings of the brain in order to

understand the behavior. We are also subject to many of the biases described in these papers, ironically. With that, it is important to consider not only the format of the information within the design of the experiment but also the type of measurement model when examining the relation between each memory ability and judgment sums. Because the interest of this project is on the unique contribution of each memory ability onto judgment sums, it is important to consider how memory is measured and quantified. Future studies on the relation between memory and judgment and decision making should consider how memory is assessed in order to account for the covariation innate to the various types of latent memory abilities. The Holy Grail will be for us to explain *why* by use of these mechanisms in order to predict the biases seen in both probability and frequency judgments.

# Appendices

## Appendix A: Experiment 1 Supplement

### Descriptives of Measures and Composites

Table 12 contains the Pearson’s correlations between each of the judgment types and distributions, the composites, and the measured variables used in experiment 1. The total number of participants, means, standard deviations, skewness, and kurtosis are also included.

Descriptives of Memory Tasks, Judgments, and Composites																	
	Memory Tasks								Judgments				Composites				
	Operation Span	Symmetry Span	Category Fluency	Experience Fluency	Delayed Free Recall	Judgment Recall Frequency Even	Judgment Recall Frequency Uneven	Judgment Recall Probability Even	Judgment Recall Probability Uneven	Frequency Judgments Even	Frequency Judgments Uneven	Probability Judgments Even	Probability Judgments Uneven	Working Memory	Retrieval Ability	Semantic Memory	Episodic Memory
Operation Span	1.000	0.396	0.256	0.201	0.423	-0.006	0.221	0.176	0.244	-0.041	-0.070	-0.047	-0.057	0.838	0.386	0.261	0.421
Symmetry Span	0.396	1.000	0.211	-0.002	0.413	0.032	0.175	0.137	0.090	0.026	0.026	-0.032	0.019	0.838	0.256	0.104	0.404
Category Fluency	0.256	0.211	1.000	0.634	0.383	0.168	0.249	0.113	0.091	-0.009	0.004	-0.238	-0.199	0.282	0.871	0.905	0.379
Experience Fluency	0.201	-0.002	0.634	1.000	0.178	0.095	0.068	0.024	0.135	0.010	0.051	-0.176	-0.158	0.113	0.779	0.905	0.166
Delayed Free Recall	0.423	0.413	0.383	0.178	1.000	0.191	0.242	0.282	0.335	-0.177	-0.201	-0.080	-0.008	0.492	0.667	0.301	1.000
Judgment Recall Frequency Even	-0.006	0.032	0.168	0.095	0.191	1.000	0.405	0.185	0.153	-0.102	-0.124	-0.252	-0.184	0.010	0.207	0.158	0.198
Judgment Recall Frequency Uneven	0.221	0.175	0.249	0.068	0.242	0.405	1.000	0.029	0.104	-0.080	-0.052	-0.154	-0.110	0.241	0.253	0.187	0.251
Judgment Recall Probability Even	0.176	0.137	0.113	0.024	0.282	0.185	0.029	1.000	0.374	-0.190	-0.200	-0.102	-0.096	0.200	0.200	0.092	0.300
Judgment Recall Probability Uneven	0.244	0.090	0.091	0.135	0.335	0.153	0.104	0.374	1.000	0.018	0.028	-0.137	-0.077	0.202	0.245	0.129	0.337
Frequency Judgments Even	-0.041	0.026	-0.009	0.010	-0.177	-0.102	-0.080	-0.190	0.018	1.000	0.930	0.183	0.183	-0.025	-0.131	-0.039	-0.232
Frequency Judgments Uneven	-0.070	0.026	0.004	0.051	-0.201	-0.124	-0.052	-0.200	0.028	0.930	1.000	0.224	0.237	-0.038	-0.132	-0.018	-0.274
Probability Judgments Even	-0.047	-0.032	-0.238	-0.176	-0.080	-0.252	-0.154	-0.102	-0.137	0.183	0.224	1.000	0.836	-0.026	-0.287	-0.306	-0.110
Probability Judgments Uneven	-0.057	0.019	-0.199	-0.138	-0.008	-0.184	-0.110	-0.096	-0.077	0.183	0.237	0.836	1.000	0.004	-0.227	-0.276	-0.026
Working Memory	0.838	0.838	0.282	0.113	0.492	0.010	0.241	0.200	0.202	-0.025	-0.038	-0.026	0.004	1.000	0.383	0.218	0.492
Retrieval Ability	0.386	0.256	0.871	0.779	0.667	0.208	0.253	0.200	0.245	-0.131	-0.132	-0.287	-0.227	0.383	1.000	0.911	0.667
Semantic Memory	0.261	0.104	0.905	0.905	0.301	0.158	0.187	0.092	0.129	-0.039	-0.018	-0.306	-0.276	0.218	0.911	1.000	0.301
Episodic Memory	0.421	0.404	0.379	0.166	1.000	0.198	0.251	0.300	0.337	-0.232	-0.274	-0.110	-0.026	0.492	0.667	0.301	1.000
N	140	141	142	140	141	141	141	141	141	142	142	142	142	137	137	137	137
Mean	24.607	15.284	21.768	23.951	12.418	3.787	3.723	3.801	3.695	110.430	98.818	131.197	125.916	0.000	0.000	0.000	0.000
SD	5.790	5.391	4.993	6.017	3.189	1.126	1.122	1.135	1.236	79.332	85.681	71.849	71.307	0.838	0.772	0.905	1.000
Skewness	-1.365	-0.594	-0.219	0.019	-0.591	-0.929	-0.689	-0.928	-0.931	3.784	4.963	2.281	2.239	-1.007	-0.473	-0.140	-0.584
Kurtosis	4.747	3.102	3.927	3.527	3.112	3.387	3.011	3.658	3.363	26.311	36.284	11.316	10.077	4.429	4.782	4.162	3.119

Table 12. Descriptives for each of the tasks and judgment variables of interest in Experiment 1.

### Impetus and Justification of Exploratory Analytic Methods

Figures 13 and 14 provide the impetus for conducting the exploratory analyses.

Figure 13 shows how each measured task relates to probability judgments. As can be seen, the three measures used to form the retrieval ability composite seem to differ in their relation to probability judgments. This was also the case for frequency judgments, which is shown in figure 14. In particular, DFR seems to have a stronger negative correlation with frequency judgments than the two fluency measures (category and experience).

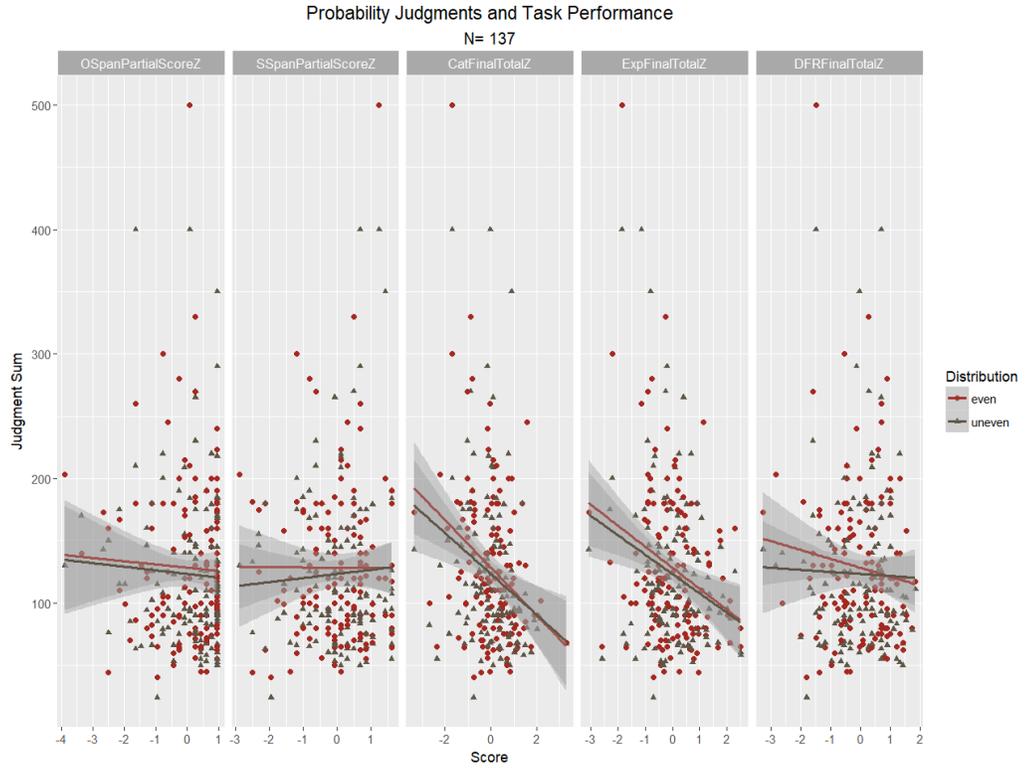


Figure 13. The relation between probability judgments and each memory measure. The grey area represents the 95% confidence interval region.

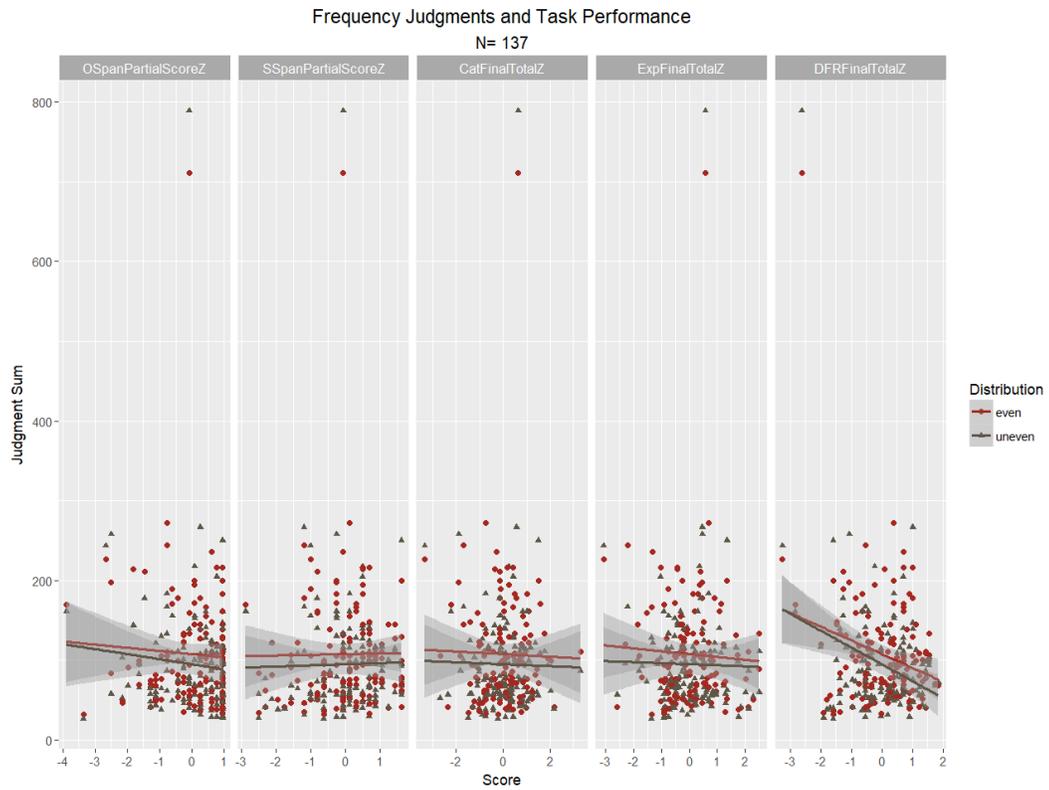


Figure 14. The relation between frequency judgments and each memory measure. The grey area represents the 95% confidence interval region.

### Common Factor Analysis vs Principal Component Analysis

Both common factor analysis (FA) and principal component analysis (PCA) describe a set of  $p$  manifest variables in terms of  $f$  latent variables and assume that  $f \leq p$ . In common FA, it is assumed that each manifest variable is a function of  $f$  common factors and one unique factor. In PCA, each manifest variable is a linear function of principal components, with no separate representation of unique variance. In other words, CFA involves reduction of the diagonal elements of the correlation matrix and PCA does not. Thus, the main difference between both methods is the parsing of the variance. Differences between the two methods will be small if the unique variances are also small (Joost C. F. de Winter & Dodou, 2014).

Further, common FA is more generalizable in that if a set of observed variables are a function of some latent ability, then a subset of those variables can also be explained by that same latent ability. Thus, factor loadings can remain consistent for different subsets of variables. PCA, on the other hand, create the components as a linear combination of manifest variables. Thus, generalizing to other sets of variables is “awkward if not possible” (J. C.F. de Winter & Dodou, 2012; J. C F de Winter, Dodou, & Wieringa, 2009; Joost C. F. de Winter & Dodou, 2014).

### Factor Analysis: Exploratory vs Confirmatory

The purpose of an EFA is to identify latent constructs or to generate hypothesis about their possible structures. The purpose of a confirmatory factor analysis (CFA) is to evaluate the hypothesized structures of the latent constructs and/or develop a better understanding of such structures (Hancock & Mueller, 2010). While we understand that it is possible to use EFA in a confirmatory manner, and additionally, it is possible

to use CFA in an exploratory manner, we found the degree of which our questions lined with EFA to be greater. Further, the study included a variant of the Category Fluency task, which was newly developed and thus, makes EFA more suitable. The findings from the EFA will be used as a guide for any subsequent and/or more confirmatory methods.

#### Extraction Method: Principal Axis Factoring vs Maximum Likelihood

Maximum Likelihood takes into account that the sample and not the population matrix is being analyzed and attempts to seek a solution that would best reproduce the population correlation values (inferential method and thus SE and GOF measures can be obtained). However, this method may not provide accurate estimates of pattern coefficients if the factors are weak and/or the sample size is too small (Briggs & MacCallum, 2003). Further, this method is not recommended if the data are not normally distributed (Fabrigar, MacCallum, Wegener, & Strahan, 1999). On the other hand, principal axis factoring uses a least squares solution which minimizes residuals between the correlation matrix being analyzed and the matrix implied by the factor model seen from the pattern coefficients and factor correlations. Principal axis factoring also is able to recover weak factors and more suitable for data that violate normality as it does not depend on this assumption.

#### Rotation Method

Because experiment 1 seeks to understand the unique contribution of retrieval ability over and beyond working memory, to the variation of judgments, when examining rotation methods, a rotation that produces uncorrelated factors is ideal. Thus,

orthogonal rotation is preferred over oblique rotation methods, which produces or allows factors to correlate. The downside of forcing orthogonality on factors is that if the structure is actually correlated, then this would cause variables to load onto more than one factor. Further, in absence of theory, oblique rotations will generally result in more reasonable representations of the data because the dimensions that underlie the constructs in the social and behavioral sciences tend to be correlated. However, our choice for orthogonal rotation lies both within our theory and our questions. Additionally, in practice, within EFA, it is acceptable to do both an orthogonal and oblique rotation and compare the results (Hancock & Mueller, 2010).

#### Exploratory Factor Analysis Supplemental Results

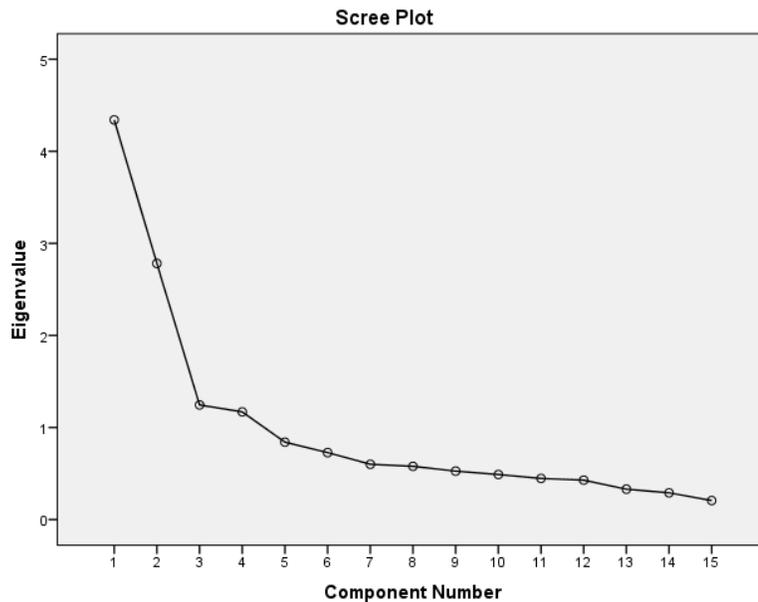


Figure 15. Scree plot of the number of components to be extracted from the EFA using Experiment 1 data.

Velicer's Minimum Average Partial (MAP) Test:

Eigenvalues  
 4.340679  
 2.781984  
 1.244603  
 1.169321  
 .840524  
 .727490  
 .599800  
 .578702  
 .525725  
 .489546  
 .446878  
 .427947  
 .330507  
 .290388  
 .205905

Velicer's Average Squared Correlations  
 .000000 .083631  
 1.000000 .057331  
 2.000000 .029000  
 3.000000 .033138  
 4.000000 .037064  
 5.000000 .044393  
 6.000000 .056928  
 7.000000 .073389  
 8.000000 .094943  
 9.000000 .125852  
 10.000000 .165724  
 11.000000 .207274  
 12.000000 .311013  
 13.000000 .483576  
 14.000000 1.000000

The smallest average squared correlation is  
 .029000

The number of components is  
 2

----- END MATRIX -----

Figure 16. Results from the MAP from the EFA using experiment 1 data.

Factor	TOTAL VARIANCE EXPLAINED											
	Initial Eigen Values			Extraction of 3 Factors			Extraction of 2 Factors			Rotation		
	Total	% of Variance	Cumulative %	Sums of Squared Loadings			Sums of Squared Loadings			Sums of Squared Loadings		
				Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.34	28.94	28.94	3.82	25.48	25.48	3.76	25.10	25.10	2.90	19.33	19.33
2	2.78	18.55	47.49	2.25	14.98	40.46	2.21	14.74	39.83	2.31	15.37	34.71
3	1.25	8.30	55.79	0.75	4.97	45.44				1.61	10.73	45.44
4	1.17	7.80	63.58									
5	0.84	5.60	69.19									
6	0.73	4.85	74.04									
7	0.60	4.00	78.04									
8	0.58	3.86	81.89									
9	0.53	3.51	85.40									
10	0.49	3.26	88.66									
11	0.45	2.98	91.64									
12	0.43	2.85	94.49									
13	0.33	2.20	96.69									
14	0.29	1.94	98.63									
15	0.21	1.37	100.00									

Table 13. Table of Eigenvalues and the total variance explained from the EFA using experiment 1 data.

FACTOR MATRICES AND SCORE COEFFICIENTS									
	Unrotated			Varimax Rotated					
	Factor Matrix			Factor Matrix			FA Score Coefficient Matrix		
	1	2	3	1	2	3	1	2	3
Ospan4	0.485	0.334	0.133	0.061	0.387	0.459	-0.005	0.048	0.137
Ospan5	0.468	0.291	0.455	0.094	0.161	0.690	0.000	-0.081	0.397
Ospan6	0.513	0.282	0.412	0.128	0.210	0.673	-0.009	-0.051	0.388
Sspan3	0.457	0.348	-0.005	0.028	0.460	0.342	-0.028	0.105	0.075
Sspan4	0.408	0.404	-0.062	-0.047	0.493	0.298	-0.056	0.133	0.064
Sspan5	0.404	0.454	0.034	-0.086	0.457	0.393	-0.040	0.116	0.116
Category Fluency1	0.682	-0.321	-0.174	0.679	0.368	0.035	0.171	0.171	-0.077
Category Fluency2	0.645	-0.463	-0.157	0.765	0.264	-0.026	0.317	0.083	-0.140
Category Fluency3	0.593	-0.381	-0.018	0.673	0.189	0.095	0.158	0.028	0.017
Experience Fluency1	0.473	-0.587	0.140	0.758	-0.084	0.083	0.308	-0.154	0.042
Experience Fluency2	0.266	-0.496	0.065	0.552	-0.122	-0.032	0.107	-0.058	0.001
Experience Fluency3	0.491	-0.439	0.168	0.657	-0.016	0.173	0.156	-0.096	0.092
DFR1	0.479	0.177	-0.110	0.170	0.452	0.199	0.016	0.099	0.003
DFR2	0.516	0.361	-0.313	0.047	0.690	0.129	-0.020	0.308	-0.085
DFR3	0.549	0.269	-0.359	0.137	0.692	0.069	-0.040	0.324	-0.114

Table 14. Factor Matrices of the unrotated and rotated solutions from the EFA using experiment 1 data.

Task	Reliability
Operation Span	0.645
Symmetry Span	0.652
Category Fluency	0.860
Experience Fluency	0.735
Delayed-Free Recall	0.625

Table 15. Spearman-Brown task reliabilities for measures used in experiment 1.

Model	Latent Construct	Variance Extracted	Reliability of Construct	Coefficient H
2-Factor	WM	0.305	0.717	0.741
	RA	0.319	0.780	0.866
3-Factor	WM	0.314	0.732	0.737
	SM	0.394	0.768	0.866
	EM	0.400	0.633	0.741

Table 16. Construct quality indices of construct quality from CFAs conducted in experiment 1.

Table 15 contains split-half reliabilities for the tasks used in experiment 1. The method used was the Spearman-Brown correction variant of the Spearman-Brown predicted reliability formula (Brown, 1910; Spearman, 1910). Note that category fluency, experience fluency, and delayed-free recall, each had 3 blocks. Thus, a pairwise correlation was done among these blocks and the average of those correlations was used in the formula. Table 16 contains different indices of construct quality. The first is the variance extracted by that construct, which is the sum of the squared standardized factor loading. The second index is the reliability of a construct, a measure of the reliability of the total score of the standardized indicators (Fornell & Larcker, 1981). Finally, the third index is Coefficient H, also referred to as maximal reliability and is the reliability of factor scores derived using regression methods (Hancock & Mueller, 2001).

### *Appendix B: Experiment 2 Supplement*

#### Descriptives of Measures and Composites

Table 13 contains the Pearson's correlations between each of the judgment types and distributions, the composites, and the measured variables used in experiment 2. The total number of participants, means, standard deviations, skewness, and kurtosis are also included.

	Memory Tasks														Judgments				Composites					
	Operation Span	Symmetry Span	Reading Span	Category Fluency	Experience Fluency	Letter Fluency	Delayed-Free Recall	Cued Recall	Gender Source	Picture Source	Gender Source Correct	Picture Source Correct	Judgment Recall Frequency	Judgment Recall Fluency	Judgment Recall Probability	Judgment Recall Fluency	Frequency	Frequency	Probability	Probability	Working Memory	Verbal Ability	Semantic Memory	Episodic Memory
Operation Span	1.000	0.508	0.345	0.167	0.091	0.167	0.297	0.232	0.111	0.206	0.029	0.084	0.180	0.088	0.107	0.175	0.053	0.036	-0.079	0.000	0.774	0.264	0.175	0.245
Symmetry Span	0.508	1.000	0.223	0.165	0.051	0.189	0.251	0.218	0.134	0.207	0.149	0.100	0.129	0.124	0.225	0.226	-0.055	-0.180	-0.081	-0.070	0.723	0.231	0.158	0.231
Reading Span	0.345	0.223	1.000	0.236	0.209	0.229	0.179	0.143	0.132	0.098	0.109	0.043	0.171	0.107	0.084	0.157	-0.025	0.013	-0.008	-0.048	0.710	0.249	0.262	0.176
Category Fluency	0.167	0.165	0.236	1.000	0.559	0.609	0.443	0.243	0.266	0.174	0.189	0.180	0.207	0.213	0.311	0.294	-0.001	-0.066	-0.144	-0.133	0.253	0.659	0.850	0.936
Experience Fluency	0.091	0.051	0.209	0.559	1.000	0.499	0.316	0.220	0.172	0.105	0.085	-0.006	0.092	0.176	0.279	0.242	0.055	0.004	-0.080	-0.086	0.159	0.544	0.817	0.724
Letter Fluency	0.167	0.189	0.229	0.609	0.499	1.000	0.559	0.261	0.251	0.158	0.119	0.026	0.183	0.164	0.188	0.016	-0.011	-0.075	-0.097	0.264	0.621	0.857	0.939	
Delayed-Free Recall	0.297	0.251	0.179	0.443	0.316	0.261	1.000	0.477	0.378	0.278	0.162	0.284	0.278	0.278	0.356	0.330	-0.014	-0.034	-0.100	-0.066	0.315	0.718	0.850	0.693
Cued Recall	0.232	0.218	0.143	0.243	0.220	0.261	0.477	1.000	0.404	0.461	0.194	0.301	0.203	0.282	0.295	0.250	-0.070	-0.084	-0.141	-0.124	0.245	0.673	0.274	0.742
Gender Source	0.111	0.134	0.132	0.266	0.172	0.251	0.278	0.404	1.000	0.589	0.069	0.155	0.055	0.164	0.153	0.150	-0.038	-0.037	-0.061	-0.021	0.159	0.552	0.267	0.542
Picture Source	0.206	0.207	0.098	0.174	0.105	0.158	0.379	0.461	0.589	1.000	0.200	0.574	0.127	0.199	0.231	0.296	0.046	0.017	-0.062	-0.112	0.245	0.634	0.173	0.754
Gender Source Correct	0.029	0.149	0.049	0.189	0.085	0.150	0.162	0.194	0.069	0.220	1.000	0.412	0.164	0.199	0.183	0.145	-0.021	-0.032	0.066	0.007	0.123	0.446	0.147	0.514
Picture Source Correct	0.084	0.108	0.043	0.180	-0.006	0.119	0.264	0.301	0.155	0.474	0.412	1.000	0.112	0.175	0.155	0.178	0.022	-0.022	-0.092	-0.105	0.104	0.551	0.112	0.600
Judgment Recall Frequency	0.188	0.129	0.171	0.207	0.092	0.106	0.259	0.303	0.055	0.127	0.164	0.112	1.000	0.447	0.111	0.104	-0.019	-0.105	-0.016	-0.033	0.164	0.235	0.150	0.276
Judgment Recall Fluency	0.088	0.124	0.107	0.213	0.176	0.183	0.278	0.262	0.144	0.199	0.199	0.175	0.447	1.000	0.216	0.181	-0.127	-0.155	-0.051	-0.079	0.125	0.356	0.244	0.311
Judgment Recall Probability	0.107	0.225	0.084	0.311	0.279	0.164	0.255	0.295	0.153	0.231	0.183	0.155	0.111	0.216	1.000	0.530	-0.105	-0.154	-0.090	-0.052	0.180	0.393	0.318	0.314
Judgment Recall Fluency	0.171	0.228	0.153	0.284	0.242	0.188	0.330	0.289	0.150	0.296	0.145	0.174	0.304	0.181	0.530	1.000	0.034	-0.031	-0.072	-0.059	0.247	0.412	0.331	0.351
Frequency	0.053	-0.055	-0.025	-0.001	0.055	0.016	-0.014	-0.028	-0.038	0.046	-0.011	0.022	-0.019	-0.127	-0.105	0.014	1.000	0.895	-0.020	-0.010	0.001	0.000	0.020	-0.013
Fluency	0.056	-0.140	0.013	-0.066	0.094	-0.011	-0.034	-0.084	-0.037	0.017	-0.102	-0.023	-0.105	-0.151	-0.154	-0.011	0.895	1.000	0.016	0.015	-0.038	-0.060	-0.036	-0.059
Probability	-0.079	-0.081	-0.080	-0.144	-0.080	-0.175	-0.103	-0.141	-0.081	-0.062	0.068	-0.092	-0.105	-0.090	-0.172	-0.020	0.016	1.000	0.888	0.018	-0.072	-0.090	-0.100	-0.076
Working Memory	0.698	-0.070	0.048	-0.131	-0.096	-0.097	-0.066	-0.124	-0.022	-0.112	0.007	-0.105	-0.023	-0.079	-0.092	-0.059	-0.016	0.015	0.888	1.000	0.901	-0.116	-0.107	-0.091
Semantic Memory	0.774	0.723	0.710	0.233	0.159	0.264	0.315	0.247	0.159	0.245	0.123	0.104	0.344	0.125	0.180	0.247	0.001	-0.028	-0.072	0.001	1.000	0.346	0.270	0.390
Episodic Memory	0.264	0.251	0.249	0.609	0.444	0.621	0.718	0.673	0.473	0.621	0.144	0.466	0.351	0.295	0.356	0.342	0.000	-0.000	-0.098	-0.116	0.345	1.000	0.728	0.961
Kurtosis	0.175	0.158	0.262	0.856	0.817	0.832	0.450	0.274	0.267	0.173	0.147	0.112	0.150	0.244	0.318	0.331	0.020	-0.036	-0.102	-0.107	0.370	0.720	0.900	0.939
Skewness	0.440	0.241	0.174	0.356	0.258	0.319	0.693	0.742	0.562	0.374	0.144	0.400	0.276	0.311	0.314	0.331	-0.013	-0.059	-0.070	-0.091	0.301	0.901	0.959	1.000
N	351	352	351	352	350	353	352	352	352	351	352	351	352	351	352	351	352	351	352	351	352	351	352	351
Mean	24.976	15.116	1.774	21.465	23.213	24.480	12.282	16.207	0.113	1.105	0.965	2.399	3.631	3.468	3.627	3.536	113.457	97.468	135.794	120.240	0.000	0.000	0.000	0.000
SD	5.780	5.176	1.851	4.216	4.745	5.746	3.453	6.067	0.896	1.146	1.321	1.201	1.815	1.221	1.210	1.200	18.876	19.843	18.163	19.149	0.745	0.600	0.851	0.663
Skewness	-1.849	-0.388	0.254	-0.190	-0.335	0.106	-0.666	0.156	-0.045	0.133	-0.484	-1.570	-0.640	-0.622	-0.876	-0.596	3.240	2.974	2.094	2.762	-0.631	-0.335	-0.153	-0.388
Kurtosis	0.439	2.548	2.440	3.578	3.071	3.083	2.875	2.709	3.150	3.034	4.091	6.014	3.170	3.041	3.000	2.893	20.810	16.213	9.131	14.423	1.163	1.722	1.150	1.866

Table 17. Descriptives for each of the tasks and judgment variables of interest in Experiment 2.

Task	Reliability
Operation Span	0.668
Symmetry Span	0.696
Reading Span	0.560
Category Fluency	0.765
Experience Fluency	0.664
Letter Fluency	0.901
Delayed-Free Recall	0.759
Cued-Recall	0.823
Picture Source Source Correct	0.855
Picture Source Correct Rejection	0.885
Gender Source Source Correct	0.653
Gender Source Correct Rejection	0.789

Table 18. Spearman-Brown task reliabilities for measures used in experiment 2.

Model	Latent Construct	Variance Extracted	Reliability of Construct	Coefficient H
2-Factor	WM	0.356	0.617	0.652
	RA	0.348	0.820	0.856
3-Factor	WM	0.328	0.590	0.609
	SM	0.551	0.785	0.798
	EM	0.323	0.718	0.793
Bi-factor	WM	0.263	0.477	0.604
	SM	0.059	0.153	0.160
	EM	0.185	0.547	0.608
2 <sup>nd</sup> Order	RA	0.222	0.744	0.824
	WM	0.365	0.617	0.704
	SM	0.157	0.354	0.364
	EM	0.320	0.714	0.790
	RA	0.506	0.737	0.869

Table 19. Construct quality indices of construct quality from CFAs conducted in experiment 2.

Table 18 contains split-half reliabilities for the tasks used in experiment 2. The same method used to calculate reliabilities in experiment 1 was used in experiment 2 tasks. Category fluency, experience fluency, letter fluency, and delayed-free recall each used the average of the pair-wise correlations between the blocks as input into the Spearman-Brown correction formula. Table 19 contains the three construct quality indices resulting from the CFAs performed on the four measurement models in experiment 2.

*Appendix C: Combined Data Analysis Supplement*

Descriptives of Measures and Composites

The table below shows contains the Pearson's correlations between each of the judgment types and distributions, the composites, and the measured variables used the combined data analysis. The total number of participants, means, standard deviations, skewness, and kurtosis are also included.

	Descriptives of Memory Tasks, Judgments, and Composites																	
	Memory Tasks									Judgments						Composites		
	Operation Span	Symmetry Span	Category Fluency	Experience Fluency	Delayed Free Recall	Judgment Recall Frequency Even	Judgment Recall Probability Even	Judgment Recall Probability Uneven	Judgment Recall Probability Uneven	Frequency Judgments Even	Frequency Judgments Uneven	Probability Judgments Even	Probability Judgments Uneven	Working Memory	Retrieval Ability	Semantic Memory	Episodic Memory	
Operation Span	1.000	0.378	0.202	0.137	0.340	0.063	0.131	0.130	0.198	0.021	-0.007	-0.066	-0.016	0.831	0.295	0.196	0.344	
Symmetry Span	0.378	1.000	0.185	0.028	0.291	0.093	0.174	0.193	0.174	-0.026	-0.072	-0.065	-0.037	0.831	0.204	0.111	0.281	
Category Fluency	0.202	0.185	1.000	0.596	0.417	0.191	0.227	0.232	0.208	-0.005	-0.033	-0.180	-0.159	0.233	0.854	0.893	0.412	
Experience Fluency	0.137	0.028	0.596	1.000	0.256	0.096	0.135	0.172	0.196	0.034	0.028	-0.119	-0.114	0.096	0.787	0.893	0.253	
Delayed Free Recall	0.340	0.291	0.417	0.256	1.000	0.216	0.267	0.332	0.333	-0.071	-0.101	-0.096	-0.043	0.376	0.709	0.373	1.000	
Judgment Recall Frequency Even	0.063	0.093	0.191	0.096	0.216	1.000	0.435	0.142	0.126	-0.050	-0.112	-0.101	-0.082	0.077	0.228	0.177	0.220	
Judgment Recall Frequency Uneven	0.131	0.144	0.227	0.135	0.267	0.435	1.000	0.159	0.158	-0.112	-0.109	-0.077	-0.085	0.156	0.282	0.216	0.278	
Judgment Recall Probability Even	0.130	0.193	0.232	0.172	0.332	0.142	0.159	1.000	0.474	-0.135	-0.170	-0.096	-0.090	0.185	0.321	0.237	0.332	
Judgment Recall Probability Uneven	0.198	0.174	0.208	0.196	0.333	0.126	0.158	0.474	1.000	0.027	0.007	-0.097	-0.064	0.218	0.333	0.249	0.339	
Frequency Judgments Even	0.021	-0.026	-0.005	0.034	-0.071	-0.050	-0.112	-0.135	0.027	1.000	0.903	0.050	0.044	-0.001	-0.039	-0.004	-0.085	
Frequency Judgments Uneven	-0.007	-0.072	-0.033	0.028	-0.101	-0.112	-0.109	-0.170	0.007	0.903	1.000	0.098	0.111	-0.045	-0.076	-0.032	-0.123	
Probability Judgments Even	-0.066	-0.065	-0.180	-0.119	-0.096	-0.101	-0.097	-0.096	-0.097	0.050	0.098	1.000	0.865	-0.069	-0.181	-0.182	-0.098	
Probability Judgments Uneven	-0.016	-0.037	-0.159	-0.114	-0.043	-0.082	-0.085	-0.090	-0.064	0.044	0.111	0.865	1.000	-0.017	-0.150	-0.172	-0.046	
Working Memory	0.831	0.831	0.233	0.096	0.376	0.077	0.156	0.185	0.218	-0.001	-0.045	-0.069	-0.017	1.000	0.500	0.185	0.376	
Retrieval Ability	0.295	0.204	0.854	0.787	0.709	0.228	0.282	0.321	0.333	-0.039	-0.076	-0.181	-0.150	0.300	1.000	0.919	0.709	
Semantic Memory	0.196	0.111	0.893	0.893	0.373	0.177	0.216	0.237	0.249	-0.004	-0.032	-0.182	-0.172	0.185	0.919	1.000	0.373	
Episodic Memory	0.344	0.281	0.412	0.253	1.000	0.220	0.278	0.332	0.339	-0.085	-0.123	-0.098	-0.046	0.376	0.709	0.373	1.000	
N	371	373	374	370	374	374	374	374	374	375	375	375	375	364	364	364	364	
Mean	24.838	15.180	21.580	23.492	12.315	3.690	3.564	3.693	3.596	112.311	97.979	134.053	122.389	0.000	0.000	0.000	0.000	
SD	5.779	5.253	4.524	5.250	3.340	1.120	1.190	1.189	1.214	81.585	76.140	77.055	69.933	0.831	0.783	0.893	1.000	
Skewness	-1.664	-0.470	-0.190	-0.089	-0.512	-0.744	-0.660	-0.902	-0.720	3.432	4.091	2.166	2.551	-0.942	-0.401	-0.196	-0.519	
Kurtosis	6.001	2.774	3.864	3.588	2.961	3.216	3.068	3.510	3.056	22.689	29.043	9.869	12.613	3.909	4.021	4.108	2.974	

Table 20. Descriptives for each of the tasks and judgment variables of interest in the combined data.

## Cross-Validation and Parameter Estimation Supplements

MCMC convergence assessments followed the initial convergence assessment procedures. This section will report the convergence of the samples used to make inference on.

*Trace, Density, & Running Means Plots.* Trace, density, and running means plots for all models were assessed visually. For all models, these plots showed convergence was good for the parameters of interest. Again, while other simulations using a different number of iterations and burn-ins were testing, the overall estimates were similar.

*Posterior Scale Reduction Factor Plots.* The posterior scale reduction factors (PSRFs) for the parameters of interest were 1. This was also the case for the multivariate posterior scale reduction factor. The PSRF shows how each parameter and the development of the scale-reduction factor over time. Evaluation of these plots suggests that each parameter appears to have reached its target distribution and appear to be fairly stable upon converging.

*Autocorrelation and Effective Sample Size.* Overall, the effective sample sizes are pretty low suggesting that there were a lot of draws that were redundant or highly correlated with other draws within a chain. Thinning was then performed in order to account for this.

*Model Diagnostics.* An examination of the residuals was done and compared to residuals that would have been obtained using maximum likelihood estimation. Additionally, the range of the Bayesian residuals and the range of the residuals fitted

using maximum likelihood were compared and across all models, these ranges were nearly identical.

## References

- Briggs, N. E., & MacCallum, R. C. (2003). Recovery of Weak Common Factors by Maximum Likelihood and Ordinary Least Squares Estimation. *Multivariate Behavioral Research*, *38*(1), 25–56.
- Brown, W. (1910). Some Experimental Results in the Correlation of Mental Abilities. *British Journal of Psychology*, *3*, 296–322. <https://doi.org/10.1111/j.2044-8295.1910.tb00207.x> LK - <https://umaryland.on.worldcat.org/oclc/5152865501>
- Conway, A. R. A., Kane, M. J., & Al, C. E. T. (2005). Working memory span tasks : A methodological review and user ' s guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786.
- de Winter, J. C. F., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, *39*(4), 695–710. <https://doi.org/10.1080/02664763.2011.610445>
- de Winter, J. C. F., & Dodou, D. (2014). Common factor analysis versus principal component analysis: a comparison of loadings by means of simulations. *Communications in Statistics - Simulation and Computation*, *0918*(December 2015), 1–39. <https://doi.org/10.1080/03610918.2013.862274>
- de Winter, J. C. F., Dodou, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*, *44*(2), 147–181. <https://doi.org/10.1080/00273170902794206>
- Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A Memory Processes Model for Judgments of Likelihood. *Psychological Review*,

106(1), 180–209.

- Dougherty, M. R. P., & Hunter, J. (2003a). Probability judgment and subadditivity : The role of working memory capacity and constraining retrieval. *Memory & Cognition*, 31(6), 968–982.
- Dougherty, M. R. P., & Hunter, J. E. (2003b). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica*, 113, 263–282. [https://doi.org/10.1016/S0001-6918\(03\)00033-7](https://doi.org/10.1016/S0001-6918(03)00033-7)
- Fabrigar, L. R., MacCallum, R. C., Wegener, D. T., & Strahan, E. J. (1999). Evaluating the Use of Exploratory Factor Analysis in Psychological Research. *Psychological Methods*, 4(3), 272–299. <https://doi.org/10.1037/1082-989X.4.3.272>
- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*. US: American Marketing Association. <https://doi.org/10.2307/3151312>
- Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, 43(2), 226–236. <https://doi.org/10.3758/s13421-014-0461-7>
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis*. Boca Raton: CRC Press. Retrieved from <http://public.eblib.com/choice/publicfullrecord.aspx?p=1438153>
- Gettys, C. F., & Fisher, S. D. (1979). Hypothesis Plausibility and Hypothesis Generation. *Organizational Behavior and Human Performance*, 24(1), 91–110.

- Gigerenzer, G., & Hoffrage, U. (1995). How to Improve Bayesian Reasoning without Instructions: Frequency Formats. *Psychology Review*, *102*(4), 684–704.  
<https://doi.org/10.1037/0033-295X.102.4.684>
- Gill, J. (2015). Bayesian Methods A Social and Behavioral Sciences Approach, 689.  
<https://doi.org/10.1198/000313008X370915>
- Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. *Structural Equation Modeling: Present and Future: A Festschrift in Honor of Karl Jöreskog*.
- Hancock, G. R., & Mueller, R. O. (2010). *The Reviewer's Guide to Quantitative Methods in the Social Sciences*. New York: Routledge. Retrieved from [http://bvbr.bib-bvb.de:8991/F?func=service&doc\\_library=BVB01&doc\\_number=018686260&line\\_number=0001&func\\_code=DB\\_RECORDS&service\\_type=MEDIA](http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&doc_number=018686260&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA)
- Hertwig, R., Pachur, T., & Kurzenhäuser, S. (2005). Judgments of risk frequencies: Tests of possible cognitive mechanisms. *Journal of Experimental Psychology: Learning Memory and Cognition*. <https://doi.org/10.1037/0278-7393.31.4.621>
- Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101.  
<https://doi.org/10.3758/BF03202365>
- Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551.  
<https://doi.org/10.1037/0033-295X.95.4.528>
- Hu, L.-T., & Bentler, P. M. (1995). Evaluating model fit. In *Structural equation*

- modeling: Concepts, issues, and applications.* (pp. 76–99). Thousand Oaks, CA, US: Sage Publications, Inc.
- Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795.
- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge: Cambridge University Press.
- Levy, R., & Mislevy, R. J. (2016). *Bayesian Psychometric Modeling*. Boca Raton: CRC Press, Taylor & Francis Group. Retrieved from <http://www.crcnetbase.com/isbn/9781439884683>
- Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged Frequency of Lethal Events. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 551–578. <https://doi.org/10.1037/0278-7393.4.6.551>
- Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York: Springer. Retrieved from <http://public.eblib.com/choice/publicfullrecord.aspx?p=602892>
- Oswald, F. L., McCabe, S. T., Redick, T. S., & Hambrick, D. Z. (2014). The development of a short domain-general measure of working memory capacity. <https://doi.org/10.3758/s13428-014-0543-2>
- Satorra, A., & Bentler, P. M. (2010). Ensuring Positiveness of the Scaled Difference Chi-Square Test Statistic. *Psychometrika*, *75*(2), 243–248. <https://doi.org/10.1007/S11336-009-9135-Y>
- Spearman, C. (1910). Correlation Calculated from Faulty Data. *British Journal of*

- Psychology*, 3, 271–295. <https://doi.org/10.1111/j.2044-8295.1910.tb00206.x>  
LK - <https://umaryland.on.worldcat.org/oclc/5152865498>
- Sprenger, A. M., & Dougherty, M. R. (2006). Differences between probability and frequency judgments : The role of individual differences in working memory capacity q. *Organizational Behavior and Human Decision Processes*, 99(2), 202–211. <https://doi.org/10.1016/j.obhdp.2005.08.002>
- Sprenger, A. M., Dougherty, M. R., Atkins, S. M., Franco-Watkins, A. M., Thomas, R. P., Lange, N., & Abbs, B. (2011). Implications of cognitive load for hypothesis generation and probability judgment. *Frontiers in Psychology*, 2(June), 1–15. <https://doi.org/10.3389/fpsyg.2011.00129>
- Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic Hypothesis Generation and Human Judgment. *Psychological Review*, 115(1), 155–185. <https://doi.org/10.1037/0033-295X.115.1.155>
- Tversky, A., & Kahneman, D. (1983). Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*, 90(4), 293–315.
- Tversky, A., & Koehler, D. J. (1994). Support Theory: A Nonextensional Representation of Subjective Probability. *Psychological Review*, 101(4), 547–567.
- Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, 37(3), 498–505. <https://doi.org/10.3758/bf03192720>
- Windschitl, P. D. (2002). Judging the Accuracy of a Likelihood Judgment: The Case

of Smoking Risk. *Journal of Behavioral Decision Making*, 15(1), 19–35.

<https://doi.org/10.1002/bdm.401>

Windschitl, P. D., & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology*, 75(6), 1411–1423.

<https://doi.org/10.1037/0022-3514.75.6.1411>

Yung, Y. F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, 64(2), 113–128. <https://doi.org/10.1007/BF02294531>