REASONING AND MEMORY: MULTIPLE SIMPLE RESPONSE STRATEGIES ARE USED IN VISUAL WORKING MEMORY FOR COLOR-ORIENTATION BINDING

A Dissertation presented to
the Faculty of the Graduate School
at the University of Missouri

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Psychology

by
KYLE O HARDMAN
Dr. Nelson Cowan, Dissertation Supervisor
MAY 2017
The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

REASONING AND MEMORY: MULTIPLE SIMPLE RESPONSE STRATEGIES ARE USED IN VISUAL WORKING MEMORY FOR COLOR-ORIENTATION BINDING

presented by Kyle O Hardman,
a candidate for the degree of Doctor of Psychology and hereby certify that, in their opinion, it is worthy of acceptance.

________________________________________
Dr. Nelson Cowan

________________________________________
Dr. Moshe Naveh-Benjamin

________________________________________
Dr. Jeffrey Rouder

________________________________________
Dr. Judith Goodman
ACKNOWLEDGMENTS

I would like to thank my advisor, Nelson Cowan, for spending the past six years helping me to develop my ability to perform research projects. I particularly appreciate how he taught me about communicating results with an audience more broad than my immediate circle of psychologists.

In addition, I would like to thank Jeff Rouder for teaching me so much about not only using statistics to solve problems, but also to think about what the statistical tests really mean. The combination of knowledge and mindset has proved invaluable to me in my research.

Thanks goes to my committee for comprehensive exams and the dissertation who have provided a number of ideas that have improved this research. Thanks goes to Moshe Naveh-Benjamin for providing the long-term memory perspective on binding. Thanks goes to Judith Goodman for providing a perspective from outside of psychology, but still very germane to my research.

A special thanks for the financial support I received for my first four years from the University of Missouri Life Sciences Fellowship program and its leaders, Dr. Mark Hannink and Debbie Allen.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS .......................................................... ii

LIST OF TABLES .............................................................. vi

LIST OF FIGURES ............................................................. vii

ABSTRACT ................................................................. xii

CHAPTER

1 Introduction ................................................................. 1

1.1 Psychometric Modeling Results ........................................ 1

1.2 Reasoning-Only Task Results .......................................... 3

1.3 Deduction Discrepancy and Hypotheses ............................... 5

1.3.1 Hypothesis 1: Maintaining Representations ..................... 5

1.3.2 Hypothesis 2: Inaccurate Models .................................. 6

1.3.3 Hypothesis 3: WM Load .......................................... 6

2 Experiment 1a .............................................................. 8

2.1 Introduction .............................................................. 8

2.1.1 Response Strategies ............................................... 9

2.1.2 Test Situations .................................................... 13

2.2 Method ................................................................. 16

2.3 Model ................................................................. 18

2.3.1 Process Trees ..................................................... 22

2.3.2 Effect of Strategy Order .......................................... 25

2.4 Results ............................................................... 27

2.4.1 Model Selection .................................................. 29
5 General Discussion .......................................................... 67
  5.1 Hypothesis 1: Maintaining Representations ..................... 67
  5.2 Hypothesis 2: Inaccurate Models ................................. 68
  5.3 Hypothesis 3: WM Load ............................................ 69
  5.4 New Hypotheses ..................................................... 70
    5.4.1 Inaccurate Metamemory ..................................... 70
    5.4.2 Degradation of Representations ............................ 71
    5.4.3 Assumption of Constant $K$ ................................ 71
    5.4.4 Partial Objects in WM ...................................... 72
    5.4.5 Motivation ................................................... 73
  5.5 Response Time Pressure ........................................... 74
  5.6 Cross-Experiment Parameter Comparisons ........................ 74
  5.7 Recall-to-Reject and Familiarity ................................ 76
  5.8 Final Conclusions ................................................ 78

APPENDIX ................................................................. 79

A Model Details ......................................................... 79
  A.1 Experiment 1a Model ............................................. 79
  A.2 Experiment 1b Model ............................................. 81

B Shrinking $\beta$ Credible Intervals ................................ 84

C Simplified Experiment 1a Model .................................... 86

REFERENCES ............................................................ 92

VITA ................................................................. 97
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Mean Parameter Values</td>
<td>76</td>
</tr>
<tr>
<td>C.1 Primary Model and Simple Model Parameter Correlations</td>
<td>90</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>19</td>
</tr>
</tbody>
</table>

2.1 Schematic of the method for Experiment 1a. Participants were shown a sample array containing 1, 3, or 4 non-hidden triangles out of a total of 6 objects (5, 3, or 2 were hidden). At test, one of three test situations was presented, where different test situations allowed some combination of Matching (M), Deduction (D), and Familiarity (F), as indicated above each test situation. At test, the participant was to select the response option (the two colored circles are response options) that the probe feature (the single, outlined triangle) was paired with in the sample array.
2.2 Binomial process trees for the Experiment 1a model. The number in the farthest-left node is the test situation that that tree applies to. The nodes labeled "M", "D", "F", and "G" indicate the start of the Matching, Deduction, Familiarity, and Guessing response strategies, respectively, with the branches following that node giving probabilities of different outcomes of that strategy. The nodes marked "A" indicate that Deduction has been attempted and that success depends on whether the necessary information is in WM. Square nodes indicate a response, where "C" indicates a correct response and "I" indicates an incorrect response. These trees are discussed further in the Process Trees subsection within the Model section of Experiment 1a.

2.3 Accuracy (left) and response latency (right) data for Experiment 1a. Error bars are 95% credible intervals. The test situations each allow for two response strategies, which are Matching (M), Deduction (D), and Familiarity (F). Test situation 1, array size 1 contains no data because a single item array cannot allow for both Matching and Deduction. The legend applies to both panels.

2.4 Posterior means and 95% credible intervals for model parameters as a function of array size for both Experiment 1a (left side of dividing line) and Experiment 1b (right side of dashed vertical dividing line). The parameters reflect WM capacity ($K$), use of Deduction ($\beta$), use of Familiarity ($\lambda$), and guessing bias toward selecting the feature response ($g$; Experiment 1b only). The $K$ parameter was constrained by the model to have the same value at all array sizes in Experiment 1a. The x-values of the Experiment 1b parameters are jittered for visibility.
3.1 Schematic of the method for Experiment 1b. The sample array contained 5 total objects, 2 of which were hidden and 3 of which were shown. At test, the participant determined whether the probe feature (the single outlined triangle) went with the feature response (colored circle), in which case the participant should click on the feature response, or not, in which case the participant should click on the square containing an X. There were five test situations that differed in terms of which response strategy could be used, indicated below the test situation (Familiarity was also possible in all test situations except for the forced guessing situation).

3.2 Binomial process trees for the Experiment 1b model. The number in the farthest-left node is the test situation that that tree applies to. Test situations 3 and 4 use the same tree. The round nodes labeled “M”, “D”, “F”, and “G” indicate the start of the Matching, Deduction, Familiarity, and Guessing response strategies, respectively, with the branches following that node giving probabilities of different outcomes of that strategy. The round nodes marked ”A”, ”A_1”, and ”A_2” indicate that Deduction has been attempted. A square node indicates a response, where ”S” indicates a Same response (the feature response is selected) and ”D” indicates a Different response (the rejection response is selected). The Same response is only correct in test situation 1.
3.3 Accuracy (left), the proportion of trials on which the feature response was chosen (center), and response latency (right) data for Experiment 1b. Error bars are 95% credible intervals. Accuracy for test situation 5 is undefined as there was no truly correct response; what is plotted is the rate at which participants were given feedback indicating a correct response.

4.1 Schematic for the Experiment 2 tasks. Top: Reasoning-only task. Bottom: Feature-matching task. The dual task is created by inserting the reasoning-only task into the maintenance interval in the feature-matching task as indicated by the dashed arrows. In the feature-matching single task, the maintenance interval is 5 s long; it is broken into three segments to show where the embedded reasoning-only task is inserted in the dual task. The reasoning-only task example shown here is a trial on which Deduction is required.

4.2 Experiment 2 accuracy and response latency for the feature-matching task (top) and reasoning-only task (bottom). For the feature-matching task, the x-axis indicates what type of reasoning was required during the maintenance interval, which was none in the single task and either Matching or Deduction in the dual task, while connect lines indicate the array size. For the reasoning-only task, the x-axis indicates secondary memory load, which was 0 in the single task and either 2 or 3 in the dual task, while connect lines indicate what kind of reasoning was required in the reasoning-only task.
C.1 Binomial process trees for the Experiment 1a simplified model. Beginning from the start node, marked S, the participant can use Matching, Deduction, or Familiarity to respond correctly. Note that each test situation prevents the use of one response strategy, which means that $F$, $D$, and $M$ are 0 in test situations 1, 2, and 3, respectively. If all memory-based response strategies fail, which happens with probability $G$, the participant enters a guessing state and responds correctly half of the time.
ABSTRACT

Recent research has examined the how effectively people use the information in working memory (WM) when reasoning about that information is required in a recognition memory task (Chen & Cowan, 2013; Hardman & Cowan, 2016). In the feature-matching task, a WM task, deductive reasoning (Deduction) can be used to determine that knowledge of one binding between colors and orientations (blue/left) precludes a different binding (blue/up), given that each color can only be paired with one other orientation. Hardman and Cowan (2016) found that the use of Deduction with the contents of WM is attempted about half of the time. In the reasoning-only task, which is very similar to the feature-matching task but lacks the requirement that items be stored in WM, however, Deduction was used accuracy and often, even by young children (Hardman & Cowan, in prep). Thus, there is a discrepancy in how people use Deduction depending on the task constraints. In this study, we sought the cause of this Deduction Discrepancy. We hypothesized that the Deduction Discrepancy may be due to (a) the requirement of WM maintenance in the feature-matching task, but not the reasoning-only task; (b) incorrect psychometric modeling results; or (c) limits the ability to use Deduction due to WM load. We examined these possibilities with novel tasks and psychometric models. Our results provided partial support for the first and third hypotheses, but not to an extent that we can fully explain the Deduction Discrepancy. New hypotheses and future directions are discussed.
Chapter 1

Introduction

Working memory (WM) research is typically focused on how many pieces of information can be held in mind (e.g. Cowan, 2001; Miller, 1956) or on the fidelity of those pieces of information (e.g. Zhang & Luck, 2008). Some research, however, has been designed to address how effectively participants are able to use the information they have in WM (e.g. Chen & Cowan, 2013; Hardman & Cowan, 2016). This is an important issue because it has implications for the accurate estimation of WM capacity (Hardman and Cowan) and because it examines the intersection between memory and reasoning. This study addresses some apparently contradictory results from two of our prior studies. We will summarize those prior results and the methods we used to obtain them, and then discuss the apparent contradiction in those results and what we will do in this study to address it.

1.1 Psychometric Modeling Results

In our first study on this topic (Hardman & Cowan, 2016), we examined to what extent participants used the information available in WM in the service of reasoning.
One of the tasks was the *feature-matching* task, which we continue to use in this study with some modifications. The basic structure of the task, however, is quite similar and can be seen in Figure 2.1. In the feature-matching task, participants were shown a number of triangles to remember, each of which had an orientation and a color. Critically, none of the colors or orientations was allowed to repeat in a given sample array, meaning that any color that went with one orientation could not also go with another orientation, and vice versa. There are many cases in the real world of items that form unique pairings, such as marriage partners, which helps to provide ecological validity for this constraint. At a delayed test, participants were required to match together the color of a triangle with its orientation, or vice versa. Participants were shown a single *probe feature* that they were required to match to one of several possible *response options*. Participants matched features (color and orientation) together at test, thus the name "feature-matching task".

In the feature-matching task, two important memory-based response strategies can be used, depending on what information is available in WM. The first response strategy can be used when the object that the probe feature was drawn from is in WM. In that case, the participant can recognize that the probe feature matches an object in WM and then select the response option that matches that object. We call this response strategy *Matching*. The second response strategy can be used when the object that the probe feature was drawn from is not in WM, but an object that one of the response options was drawn from is in WM. In this case, participants are unable to use Matching. They are, however, able to reject a response option as incorrect by the following logic: "I know that this response option was part of an object that had a different feature than the probe feature and, because colors and orientations cannot repeat, I know the probe feature must have gone with a different response option.” This response strategy is a kind of deductive reasoning, so we call it *Deduction*.

Using a psychometric model, Hardman and Cowan (2016) estimated both WM
capacity and how often participants used Deduction when they were able to do so. Those authors found that participants used Deduction about 60% of the time. This result was approximately consistent with the results of an additional change-detection task also used by Hardman and Cowan, in which participants appeared to use Deduction about 50% of the time. Thus, Hardman and Cowan found converging evidence that their average participant used Deduction in moderation. This result on its own is plausible, but apparent contradictions arose in the results from a follow-up study.

1.2 Reasoning-Only Task Results

In a follow-up study, (Hardman & Cowan, in prep) examined, among other things, to what extent participants used Matching and Deduction when they were not put under a WM load. The task we used was a variant of the feature-matching task in which the sample array items were shown on screen while participants were being tested. Thus, participants were not required to maintain items in WM across a delay, which could affect their use of response strategies, or at least what information was available to them. We call this task the reasoning-only task because it only requires reasoning about the test display, but not maintenance of items in WM. We only used two response options so that successful use of Deduction in eliminating the incorrect response option would guarantee a correct response.

The reasoning-only task also differs from the feature-matching task in that it is possible to control what response strategy participants could use. This is accomplished by using sample arrays in which only some of the triangles (most often only one triangle) are shown to participants. The premise is that the participant was to imagine that they had seen a whole sample array but only remembered a few of the items. The benefit of using partial sample arrays is that it forces participants to use the response strategy of the experimenter’s choice. To force the use of Matching, we
would present a sample array with an object that the probe feature and one of the response options would be drawn from. The other response option would be drawn from an object that was not shown, preventing the use of Deduction with that response option. Thus, participants could only use Matching to respond. To force the use of Deduction, we would present a sample array with an object that one of the response options would be drawn from, but not the object that the probe feature would be drawn from. A participant can use Deduction by recognizing that the known response option went with a different feature than the probe feature, thus rejecting that response option. We told participants, truthfully, that we always gave them enough information to respond correctly.

When participants were forced to use Deduction, Hardman and Cowan (in prep) found that participants were equally accurate or more accurate than when participants were forced to use Matching.\textsuperscript{1} Notably, this study used a cross-sectional developmental design with child participants of various ages. Even for our youngest group of participants, who were 6 to 7 years old, accuracy on Deduction trials was equivalent to accuracy on Matching trials. For that age group, accuracy on Deduction trials was 96%, which seems to be quite good for participants of that age. Older age groups showed steady increases, up to 99% accuracy for adults. In addition, to anticipate our findings, we fundamentally replicate this result in adult participants in Experiment 2 of this study. These results suggest that, provided with the right information, participants are very good at using Deduction, even from a young age. This suggests that Deduction is fairly easy to do, which implies that Deduction might be used by participants in the feature-matching task at high rates.

\textsuperscript{1}The greater accuracy for Deduction than Matching could be due to the reward structure used by Hardman and Cowan (in prep): Participants were given more points for a correct Deduction response than a correct Matching response. This could have motivated them to focus on the Deduction strategy rather than the Matching strategy. In Experiment 2, we used equal reward for Matching and Deduction trials in the reasoning-only task and found slightly worse performance on Deduction trials than Matching trials.
1.3 Deduction Discrepancy and Hypotheses

Now that we have summarized the results of Hardman and Cowan (2016) and Hardman and Cowan (in prep), we can explain why those results appear to be contradictory. First, Hardman and Cowan (2016) found that in the feature-matching task, participants appear to accurately use Deduction about 60% of the time. Then Hardman and Cowan (in prep) found that in the reasoning-only task, participants are able to accurately use Deduction nearly all of the time, even from a young age. If, as is suggested by the results from the reasoning-only task, Deduction is fairly easy to do, it is not clear why participants only use Deduction only 60% of the time in the feature-matching task. We will refer to this as the Deduction Discrepancy. In this study, we will investigate two major hypotheses about why the results from our past studies appear to be in conflict.

1.3.1 Hypothesis 1: Maintaining Representations

Our first hypothesis is that the Deduction Discrepancy is due to the fact that the stimuli are always present on screen in the reasoning-only task, while those stimuli must be maintained in WM during the feature-matching task. Perhaps in the feature-matching task the quality of the representations in WM are of lower quality after maintenance, which somehow limits the usefulness to Deduction but still allows those representations to be used for Matching. Regardless for the reason why, it is possible to test this hypothesis by using a version of the reasoning-only task in which there is a delay between study and test. We do this in Experiment 1a. This hypothesis predicts that the Deduction Discrepancy would appear if the items needed for the reasoning-only task are maintained in WM. Thus, evidence in favor of this hypothesis would consist of substantially reduced accuracy on Deduction trials relative to Matching trials on the delayed reasoning-only task.
1.3.2 Hypothesis 2: Inaccurate Models

Our second hypothesis is that participants actually use Deduction nearly all of the time regardless of whether they are doing the feature-matching or reasoning-only tasks, but the psychometric modeling results from the feature-matching task were inaccurate. The model for the feature-matching task used by Hardman and Cowan (2016) is limited by the data that were provided to it. In particular, on any given trial, there was no control over the response strategies used by participants, so either Matching or Deduction could have been used. Although the model was theoretically capable of separating out Deduction from Matching, a model that had more information about which response strategies were available to participants would be able to more accurately estimate how often Deduction is used. In Experiments 1a and 1b, we use new tasks that give us more control over the response strategies available to participants. This control over response strategies is included in the data given to the new models we develop, which allows those models to more accurately estimate Deduction abilities. This hypothesis would be supported if the results of the new tasks and models are substantially different from the results of Hardman and Cowan. If we instead replicate the result of Hardman and Cowan, this hypothesis would be rejected.

1.3.3 Hypothesis 3: WM Load

Our third hypothesis is that the Deduction Discrepancy is due to the fact that WM is loaded with objects at test, but using Deduction requires at least some of the WM resources taken up by those objects. One possibility is that Deduction requires WM resources for intermediate computations, as suggested by Barrouillet and Lecas (1999). If WM were fully loaded, a participant would have to choose between not using Deduction or dropping some items from WM so that Deduction could be used.
The resource competition could result in either or both of Matching or Deduction being hurt, but if the impact on Deduction is larger, that could lead to the Deduction Discrepancy. We expand on these ideas and examine this hypothesis in Experiment 2. In that experiment, we had participants perform the reasoning-only task while under a secondary WM load. We include a no-WM-load condition as a baseline, which serves as a replication of the reasoning-only task results of Hardman and Cowan (in prep). In order to support this hypothesis, we must find evidence that Deduction is hurt more than Matching by the WM load manipulation. Thus, this hypothesis can be supported if we find an interaction between WM load and response strategy in which Deduction is selectively hurt by WM load.
Chapter 2

Experiment 1a

2.1 Introduction

In Experiments 1a and 1b, we address Hypothesis 2: Inaccurate Models by better measuring to what extent participants are using Deduction than we have in past studies (Hardman & Cowan, 2016). In Experiments 1a and 1b, we use modified versions of the feature-matching task. The fundamental nature of the modification is to create conditions in which participants cannot use Matching but instead are compelled to use Deduction or another response strategy. The critical modification is to show participants only some of the objects in the sample array, replacing the hidden objects with placeholders that indicate only the presence of an object but not its identity. This manipulation simulates WM failure: The hidden items are like items that were never encoded into WM or were forgotten during maintenance. As in the reasoning-only task, hiding some objects in the sample array allows the experimenter to put participants into test situations in which the participant is prevented from using certain response strategies.

As one example, it is possible to put participants in test situations in which the
participant was never shown the object that the probe feature was drawn from. This lack of information prevents the participant from using Matching because they cannot remember the pairing between the probe feature and the correct response option if they were never shown the relevant object. When the use of Matching is prevented, it is still possible to allow for alternative response strategies, such as Deduction.

The value of these designs is that it gives us information about which response strategies were or were not available to participants in the different conditions. This allows for the construction of models that take advantage of this additional information. Information about the possible response strategies was not available to the model used by Hardman and Cowan (2016), which limited the ability of that model to precisely estimate Deduction usage.

This experiment also addresses Hypothesis 1: Maintaining Representations. We include a condition in which participants are effectively performing the reasoning-only task, but with a maintenance interval between study and test. We accomplish this by including trials in which only one object is shown to participants and the other objects are hidden, as was common in the reasoning-only task, and by using only two response options, as Hardman and Cowan (in prep) did. If performance is high when only one object is shown to participants in this task, it suggests that WM maintenance alone is not enough to account for the Deduction Discrepancy.

### 2.1.1 Response Strategies

The response strategy that we are primarily interested in is Deduction, but there are four total response strategies that we will examine in Experiments 1a and 1b. The first two response strategies are Matching and Deduction, which we have discussed. We also consider uninformed guessing, in which participants have no information, to be a response strategy. We do not focus on it, but do account for it in our models. Finally, in our tasks, it is possible for participants to use their familiarity with individual
features (colors and orientations) to respond with above-chance accuracy even without knowledge of the bindings between features that create whole objects. It is difficult to prevent participants from using familiarity, so we account for the possibility that they do so in the models. We have discussed Matching and Deduction in the Introduction of this study. We will now discuss uninformed guessing and familiarity further.

**Uninformed Guessing**  Uninformed guessing is the response strategy used by participants when they have no information about the correct answer. This might happen because the participant failed to attend to the sample array and has no information about any objects on the trial. It can also happen when the objects that a participant remembers are all different from the objects that the participant is tested on. Hardman and Cowan (2016) treated ideal information use in the feature-matching task as a kind of informed guessing, in which Deduction could be used to exclude incorrect response options, which was followed by the participant guessing from among the remaining response options. This treatment made sense given that the test situation used by Hardman and Cowan included all possible response options, so the exclusion of some of the response options would not guarantee a correct response. In this study, we instead use test situations with two possible response options, so by using Deduction to exclude one response option, the participant is able to respond correctly without any guessing. Thus, the only kind of guessing we consider here is uninformed guessing as no informed guessing is possible in these experiments.

**Familiarity**  When making decisions about which response to make, participants may be able to use a kind of familiarity-based reasoning (which we shorten to Familiarity) to guide them to the correct answer, at least with above-chance success rates. It should be noted that our use of "Familiarity " differs from how that word is used in long-term memory research, which is discussed further in the Recall-to-Reject
and Familiarity section in the General Discussion. An important part of the ability to use Familiarity rests on the assumption, on the part of participants, that whole objects are stored in WM, such that if one feature of an object is known, the other feature of that object will be known as well. To be clear, we are not endorsing this view and even have clear evidence against it (Hardman & Cowan, 2015; Oberauer & Eichenberger, 2013). Even if it is not necessarily the case that whole objects are always stored in WM, it could still be possible that knowledge of features from objects are selected from the sample array in such a way that if one feature of an object is known, it is more likely that other feature of the object is known (positively correlated feature knowledge). In a design with two-feature objects, Fougnie and Alvarez (2011) concluded that feature knowledge was more independent than positively correlated, but allowed for some positive correlation between features of objects. Regardless of whether participants are conversant in the WM literature, it is quite plausible that they might believe that knowledge of one feature of an object necessitates knowledge of the other feature, or is at least that knowledge of features is positively correlated.

Either of these beliefs about how features of objects are stored in WM allows for the use of Familiarity to guide responding. One case in which Familiarity can be used is when a participant is familiar with the probe feature and one response option, but unfamiliar with the other response option. For example, in test situation 2 in Figure 2.1, the participant remembers the color green and orientation top-left, but not that they were paired, and does not remember purple. If whole objects are stored in WM, it seems implausible that a familiar and an unfamiliar feature could have gone together. If those features had gone together and the participant was familiar with one of the features, the reasoning goes the participant should also be familiar with the other feature. Thus, if the probe feature is familiar and only one of the response options is unfamiliar, the participant should select the familiar response option. This basic logic also works in a probabilistic way if knowledge of the features of objects is
positively correlated. If the probe feature and only one response option are familiar, then the participant believes that it is likely that the familiar response option was paired with the probe feature and therefore selects the familiar response option.

Another case in which Familiarity can be used is when the probe feature and one response option are unfamiliar, but the other response option is familiar. In this case, the familiar response option and the unfamiliar probe feature probably did not go together, otherwise both would be familiar. As a result, the familiar response option is rejected and the unfamiliar response option is selected. According to this way of thinking, the unfamiliar probe feature (e.g., the orientation left in test situation 3 in Figure 2.1) and unfamiliar response option (e.g., the color purple) are assumed to come both from hidden objects or at least objects that are not in working memory, and therefore more likely to go together than an unfamiliar probe and a familiar color response.

According to the preceding logic, it is possible for participants to use familiarity cues in some cases to guide responding. Importantly, Familiarity does not require information about the binding between features, simply knowledge of (or familiarity with) the features themselves. In Experiments 1a and 1b, some of the objects in the sample array are hidden, which means that we know that those objects must be unfamiliar to participants. Non-hidden objects may or may not be familiar to participants, depending on whether they remember those objects. The potential for participants to use Familiarity means that we should account for it in our models and the designs of the experiments allow us to. Although we account for Familiarity, we treat it as a nuisance response strategy of little interest to our primary goal of estimating how often Deduction is used.
2.1.2 Test Situations

In this experiment, there are three test situations, each of which has different possible response strategies. Refer to Figure 2.1 throughout this section for a depiction of the test situations. In the sample array shown in that figure, there are two presented objects that are important for the following descriptions: the green object pointing to the top-right and the orange object pointing to the top-left. In addition, there is one important hidden object, which was purple and pointed left.

In test situation 1, participants are able to use Matching and Deduction. This is accomplished by (a) drawing the probe feature and one of the response options from the same non-hidden object in the sample array (the green object in the figure) and (b) drawing the other response option from another non-hidden object in the sample array (the orange object). Participants may remember the object from which the probe feature was drawn, which would allow them to use Matching to pair together the probe feature with the matching response option. Alternatively, they can use Deduction to recognize that the non-matching response option was originally paired with something other than the probe feature (orange paired with top-left, not top-right). This use of Deduction would allow them to eliminate the orange response option from consideration and select the correct response. Note that this condition is very similar to the feature-matching task test situation used by Hardman and Cowan (2016) and Hardman and Cowan (in prep). The key innovation in this experiment is the inclusion of the other two test situations.

In test situation 2, participants are able to use Matching and Familiarity. This is accomplished by (a) drawing the probe feature and one of the response options from the same non-hidden object in the sample array (the green object) and (b) drawing the other response option from one of the hidden objects in the sample array (the purple object). As in test situation 1, participants may remember the object from which the probe feature was drawn, which would allow them to use Matching
to match together the probe feature with the matching response option (top-right with the green response option). Alternatively, participants might use Familiarity to recognize that the probe feature (top-right) and one response option (green) are both familiar, but that the response option from the hidden object is not familiar. Because the probe feature and only one response option is familiar, the participant should select the familiar response option, which would be the correct response in this test situation.

In test situation 3, participants are able to use Deduction and Familiarity. This is accomplished by (a) drawing the probe feature and one of the response options from one of the hidden objects in the sample array (the purple object) and (b) drawing the other response option from one of the non-hidden objects in the sample array (the orange object). Participants may remember the object from which the non-hidden response option was drawn, which would allow them to use Deduction to recognize that the non-hidden response option did not go with the probe feature. Alternatively, participants might use Familiarity to recognize that the probe feature (left) and one response option (purple) are both unfamiliar, but that the other response option (orange) is familiar. By the logic of Familiarity, if orange went with left, then both orange and left should be familiar. Since only orange is familiar, then the unfamiliar left probe feature must not have gone with orange and the participant should select the unfamiliar purple response option, which is the correct response in this test situation.

Although we have described test situation 1 as allowing the Matching and Deduction response strategies, it is possible for participants to use Familiarity in test situation 1. For example, a participant might be familiar with both features of the object that the probe feature was drawn from, but not remember that those features were necessarily paired together in the sample array. In this case, they would be able to successfully use Familiarity. This is only one possible state of knowledge, however, and there are a total of eight possible familiarity states with the probe feature and the
two response options. When these eight states are enumerated, Familiarity cannot be used in four of them, leads to the incorrect response in two of them, and leads to the correct response in two of them. In order to fully model Familiarity, we would have to be able to estimate how often each of the eight knowledge states occur. Modeling the full complexity of this situation is impossible without having dramatically more rich data than we have available. Thus, we must make a simplifying assumption to operationally define Familiarity.

There is one simplifying assumption that allows us to still estimate Familiarity, albeit in an inexact way. This assumption is that all of the features in the test display that were also presented in the sample array are familiar to participants when Familiarity is used. One argument in favor of this assumption is to think that even a weak familiarity signal could be enough to use Familiarity. Features that were presented on a given trial could provide enough of a familiarity difference when compared with features that were not presented on that trial to allow for their relative familiarity to be used.

This simplifying assumption makes it so that participants are unable to use Familiarity in test situation 1 because all of the features in the test display would be familiar. In test situations 2 and 3, this assumption allows participants to make the correct response if they use Familiarity, which is consistent with the model we have described. Thus, assuming full familiarity with the test display features (at least those that were present in the sample array) allows for a simple operational definition of Familiarity. Even if this definition of Familiarity is inexact, it should be better than not modeling Familiarity at all and should improve the quality of our estimates of the parameters that we are primarily interested in. This definition is also used in Experiment 1b.
2.2 Method

There were a total of 31 participants (16 female; mean age 18.7 years) drawn from an introductory psychology course at the University of Missouri. Participants were given partial course credit for participating. All participants were treated in accordance with the ethics requirements of the University of Missouri institutional review board.

The structure of the trials of the task are shown in Figure 2.1. Each trial was initiated by the participant by clicking the mouse anywhere on the screen. Once the trial was initiated, there was a 500 ms black screen which was followed by the sample array, which was presented for 1000 ms. Following the sample array, there was a 1000 ms maintenance interval during which the screen was blank. Following maintenance, the test display was displayed until a response was made. The participant used the mouse to click on the response option that matched the probe feature. The probe feature and response options were individual features, which were created by removing one of the features from a triangle. A color without orientation information was created by using a colored circle (a circle points in all directions equally). An orientation without color information was created by using an outlined triangle with the outline drawn in white (white was not one of the colors used for non-hidden triangles in the sample array).

When the participant hovered the mouse over a response option, an arrow pointing from the probe feature to the response option was shown. After a response was made, the arrow remained for the feedback display. When a correct response was made, a green check mark was drawn over the arrow from the probe feature and the selected response option. When an incorrect response was made, a red X was drawn over the arrow instead. In addition, participants were awarded points for correct responses that were tracked over the course of the experiment. As can be seen in the feedback display in Figure 2.1, points were shown above the probe feature as a simple equation. The previous score was shown on the left, the points earned from the current trial
were shown in the middle, and the sum of the previous and current points was shown on the right. If the response was correct, 40 points were earned and newly-earned points were drawn in a green font. If the response was incorrect, 0 points were earned and the 0 was drawn in a red font.

The sample arrays always contained 6 total objects, of which some number were “hidden”: replaced with a circle with a question mark in it. On different trials, either 1, 3, or 4, objects were not hidden of the 6 total objects. The number of non-hidden objects defines the levels of the array size factor. From the perspective of the participant, although there are hidden objects presented, there is no information to be gained from those objects, so only the non-hidden objects should be encoded into WM, perhaps along with the number of hidden objects in the array or other information of secondary importance. Each object in the sample array was made by combining one of the eight possible colors with one of the eight possible orientations. Colors and orientations were sampled without replacement, which resulting in each color and orientation being unique in a given sample array. The uniqueness of colors and orientations is required for participants to be able to use Deduction: If colors and orientations repeated, knowing that, e.g., red went with left would not allow a participant to determine that red did not go with down.

The possible colors of the objects follow, with the RGB coordinates of those colors in parentheses: red (229, 0, 0), orange (229, 140, 0), yellow (255, 229, 0), green (0, 229, 57), cyan (0, 255, 255), blue (25, 82, 255), purple (130, 26, 242), and pink (229, 0, 172). The colors were chosen to be perceptually discriminable. The background color was a dark grey (40, 40, 40). All text, the question mark stimuli, the response arrow, and the outlines of the outlined triangles were all white (255, 255, 255). The colors were the same in all of our experiments. The orientations were separated by 45° of angle, starting at 0°.

Participants were given instructions about how to do the basic feature-matching
task with no hidden objects in the sample array. We wanted participants to be aware of how to use Deduction on the task, so they were given explicit instructions about how to do so. These included images showing an example trial and verbal instructions on using Deduction to correctly answer in the example. Following basic instructions, participants performed 16 practice trials of the basic feature-matching task. Then participants were given instructions about the main task, with hidden objects in the sample array, and were left to finish the experiment. Participants performed two trials blocks of 200 trials each, for a total of 400 trials. These trials were approximately evenly divided between the 8 array size by test situation cells (array size 1, test situation 1 being excluded), for approximately 50 trials per cell. Trials from each cell were randomly intermixed in the two trial blocks.

The computer programs for all of the experiments in this study were created with CX (Hardman, 2016). The experiments were run on computers with 15-inch CRT monitors running at a resolution of 1024x768 pixels. Participants used computer mice to respond; these mice were standard two-button mice. Participants were left unobserved, seated in sound-attenuated booths for the duration of their experimental sessions, which lasted approximately an hour.

2.3 Model

Our model accounts for the four response strategies that participants are able to use in this task. Participants either attempt the response strategies in some order or attempt all simultaneously, but we do not know which. We will assume a sequential ordering for reasons of modeling simplicity. Also for the sake of simplicity, we will make an assumption about the order in which participants attempt to use the response strategies. In particular, we will assume that participants use Matching, Deduction, Familiarity, and uninformed guessing in that order. These assumptions should not
Figure 2.1: Schematic of the method for Experiment 1a. Participants were shown a sample array containing 1, 3, or 4 non-hidden triangles out of a total of 6 objects (5, 3, or 2 were hidden). At test, one of three test situations was presented, where different test situations allowed some combination of Matching (M), Deduction (D), and Familiarity (F), as indicated above each test situation. At test, the participant was to select the response option (the two colored circles are response options) that the probe feature (the single, outlined triangle) was paired with in the sample array.
have a substantial effect on the modeling results, although we do acknowledge that these assumptions, if false, will negatively impact the validity of the results to some extent. After discussing the model in more detail, we explain in the Effect of Strategy Order subsection how the model would change if we made different assumptions about response strategy usage order. As we will show, the model would change only minimally if different assumptions were made.

The model contains a number of parameters that are related to WM capacity, the various response strategies, and an inattention. We describe those parameters here, including their interpretation, before moving on to their placement in the model.

WM capacity is estimated with the $K$ parameter. This parameter estimates is the average number of objects for which their color, orientation, and the binding between color are orientation are known. Both the Matching and Deduction response strategies depend on $K$. We assume that Matching is always attempted first and can only fail due to lack of knowledge. We make these assumptions because Matching is a fairly simple response strategy that is unlikely to fail for reasons other than lack of knowledge. Thus, if the $K$ objects in WM do not contain the object from which the probe feature was drawn, Matching fails. The success of Deduction also depends on $K$ because Deduction can only be performed if the necessary object is in WM.

The $\beta$ parameter reflects the probability that a participant attempts to use Deduction. Once an attempt to use Deduction is made, it is still possible to fail to successfully use Deduction if the required information is not in WM, which depends on the $K$ parameter. If the participant has the required information, they are able to answer correctly; if they do not, they attempt a different response strategy. It is possible that failures of Deduction happen even if the required information is available in WM. Thus, strictly speaking, $\beta$ does not reflect attempts to use Deduction, but attempts that are successful given that the necessary information is in WM.

The parameter related to Familiarity, $\lambda$, is the probability that a participant
succeeds at using Familiarity. Once a participant attempts to use Familiarity, they can fail if the required information is not available. We cannot, however, model the information required for Familiarity, namely individual features of objects (it takes specialized tasks to model feature memory). Thus, we cannot determine how often Familiarity would fail once attempted, so $\lambda$ reflects how often Familiarity is successful. Although Deduction and Familiarity can fail due to the necessary information not being in WM, we model the information needed for Deduction with the $K$ parameter, but do not model the information needed for Familiarity. In summary, we model Deduction failures due to having insufficient information, but not Familiarity failures for the same reason, which results in $\beta$ and $\lambda$ having subtly different interpretations.

There are no parameters specifically related to guessing. Guessing occurs when all other strategies fail, so the proportion of trials on which a participant must guess depends on other parameters of the model. When a participant guesses, they have no information about the correct answer, so their probability of guessing correctly depends only on the number of response options. In this experiment, there are two response options and one response is always correct, so the probability of guessing correctly is $1/2$.

We assume that participants sometimes fail to correctly perform the task due to inattention. Inattention captures a variety of errors, such as failing to encode the sample array or making a response other than the intended response. Including an attention parameter is important in fixed-capacity models in order to account for errors at low array sizes (Rouder, Morey, Morey, & Cowan, 2011). In test situation 2 at array size 1, which we used in this experiment, if full attention is assumed and the model estimates that a participant has more than 1 item in WM on average (i.e. $K > 1$), then the model will predict that the participant will respond correctly 100% of the time. Perfect performance is, however, very uncommon in any task no matter how simple, so any errors at low array sizes would result in $K$ estimates to be biased.
low. We were unable to estimate the attention parameter in the model as model fit diagnostics indicated that the attention parameter was not identifiable given the other parameters of the model. Rather than dropping the attention parameter, we fixed it to the constant value 0.95, which means that inattention was assumed to happen 5% of the time. This value was based on average accuracy in test situation 2 at array size 1, which was 95.3%. When participants suffer from an attention lapse, the model assumes that they have no information and must guess, guessing correctly with probability 1/2.

We estimated the parameters of our model using Bayesian methods which are discussed further in Appendix A.

2.3.1 Process Trees

Figure 2.2 shows binomial process trees for each of the three test situations. The leftmost node in each tree contains a number indicating the test situation to which the tree corresponds. The nodes labeled ”M”, ”D”, and ”F” indicate decision points related to the Matching, Deduction, and Familiarity response strategies, respectively, while the G node is related to uninformed guessing. For example, at an Matching node, a participant either will or will not have in WM the information needed to perform Matching. If the information is available, the appropriate branch is taken and a correct response is made. If the information is not available, the other branch is taken and a different response strategy is attempted. Thus, the Matching, Deduction, and Familiarity nodes should be thought of as the start of those response strategies and the following branches as probabilities of different outcomes of that strategy.

In Figure 2.2, $K$ is the number of objects in WM, $N$ is the number of non-hidden objects in the sample array, $\beta$ is the probability that the participant will use Deduction, and $\lambda$ is the probability that the participant will successfully use Familiarity. Given the nature of the model, reaching a node in which the participant will attempt
to use Deduction or Familiarity is conditional on other, previous branches of the model (at least in some trees). Thus, $\beta$ and $\lambda$ should be interpreted as probabilities conditional on being in a position to attempt to use Deduction or Familiarity, respectively, rather than as proportions of trials.

With respect to the model trees in Figure 2.2, the attention step happens before the leftmost node in the trees. For each test situation, with probability 0.95, the participant was attentive and reaches the leftmost node in the tree and can use memory-based strategies. With probability 0.05, the participant was inattentive and must guess, guessing correctly with probability 1/2.

**Test Situation 1** In test situation 1, the participant is able to use Matching and Deduction. The model assumes that the participant first attempts to use Matching, followed by Deduction if Matching fails. This ordering can be seen in the binomial process tree for test situation 1 shown in Figure 2.2. Beginning at the start node, marked ”1”, they move to the ”M” node and attempt to use Matching. With probability $\min(K/N, 1)$, the participant knows the object that allows for Matching to occur. The upper bound of 1 is used for $K/N$ to account for the fact that we will estimate $K$ based on all array sizes, but at smaller array sizes, $K$ might be larger than $N$ and it is not possible to respond correctly with a probability greater than 1. In the example shown in Figure 2.1, if the participant knows the green triangle from the sample array, they are able to use Matching. If Matching fails due to lack of knowledge, the participant moves to the node marked ”D”. From the ”D” node, the participant attempts to use Deduction with probability $\beta$, in which case they move to the Deduction attempted node, marked ”A”. The probability of having the information in WM that would allow for the successful use of Deduction is $\min(K/(N-1), 1)$

1. If the necessary information is in WM, which is the orange object in the exam-
ple shown in Figure 2.1, the participant will respond correctly. The probability of successfully using Deduction is \( \beta \ast \frac{K}{N - 1} \). If the participant either does not attempt to use Deduction or fails to use Deduction because they lack the necessary information, they have exhausted both memory based response strategies and must guess. From the guessing node marked "G", the participant guesses correctly with probability 1/2.

**Test Situation 2** In test situation 2, the participant is able to use Matching and Familiarity. The model assumes that participants first attempt to use Matching, followed by Familiarity if Matching fails. This ordering can be seen in the binomial process tree for test situation 2 shown in Figure 2.2. Beginning at the start node, marked "2", they move to the "M" node and attempt to use Matching. With probability \( \min(K/N, 1) \), the participant knows the object that allows for Matching to occur. If Matching fails due to lack of knowledge, the participant moves to the node marked "F". From the "F" node, with probability \( \lambda \) the participant successfully uses Familiarity and responds correctly. If Familiarity fails, the participant has exhausted both memory based response strategies and must guess. From the guessing node marked "G", the participant guesses correctly with probability 1/2.

**Test Situation 3** In test situation 3, the participant is able to use Deduction and Familiarity. The model assumes that participants first attempt to use Deduction, followed by Familiarity if Deduction fails or is not attempted. This ordering can be seen in the binomial process tree for test situation 3 shown in Figure 2.2. Beginning
at the start node, marked ”3”, the participant moves to the ”D” node and attempts to use Deduction with probability $\beta$, in which case they move to the Deduction attempted node, marked ”A”. The probability of having the information in WM that would allow for the successful use of Deduction is $\min(K/N, 1)$. Note that the denominator is $N$, rather than $N - 1$ as in test situation 1, because an attempt to use Deduction in this test situation is not conditional on failing to use Matching due to having insufficient knowledge. If the information necessary for Deduction is in WM, the participant will respond correctly. If the participant either does not attempt to use Deduction or fails to use Deduction because they lack the necessary information, the participant moves to the node marked ”F”. From the ”F” node, with probability $\lambda$ the participant successfully uses Familiarity and responds correctly. If Familiarity fails, the participant has exhausted both memory based response strategies and must guess. From the guessing node marked ”G”, the participant guesses correctly with probability $1/2$.

### 2.3.2 Effect of Strategy Order

Now that we have laid out the model, let us return to the assumption about the order in which response strategies are used. We will make the argument that using different orders has little effect on the model. We begin with test situation 2. In that test situation, the probability of making an incorrect response (given that attention was paid) is $(1 - K/N)(1 - \lambda)(1/2)$ (ignoring the min on $K/N$). That probability is based on the assumption that Matching is used before Familiarity. If we instead assume that Familiarity is used before Matching, the probability is instead $(1 - \lambda)(1 - K/N)(1/2)$, which evaluates to the same value. Thus, for test situation 2, assuming a different order of response strategies has no effect. The same is true of test situation 3. For test situation 1, however, there is a small difference. Assuming Matching before Deduction, the probability of an incorrect response is $(1 - K/N)(1 - \beta K/(N - 1))(1/2)$. 

25
Figure 2.2: Binomial process trees for the Experiment 1a model. The number in the farthest-left node is the test situation that that tree applies to. The nodes labeled "M", "D", "F", and "G" indicate the start of the Matching, Deduction, Familiarity, and Guessing response strategies, respectively, with the branches following that node giving probabilities of different outcomes of that strategy. The nodes marked "A" indicate that Deduction has been attempted and that success depends on whether the necessary information is in WM. Square nodes indicate a response, where "C" indicates a correct response and "I" indicates an incorrect response. These trees are discussed further in the Process Trees subsection within the Model section of Experiment 1a.
Assuming Deduction before Matching, the probability of an incorrect response is 
\((1 - \beta K/N)(1 - K/(N - 1))(1/2)\). The difference is whether the denominator for \(K\)
is \(N\) or \((N - 1)\). We do not expect this difference to substantially affect the results,so we are comfortable assuming the order of response strategies.

2.4 Results

All of our analyses are Bayesian in nature, typically using Bayes factors to compare hypotheses. In each hypothesis test, there is a null hypothesis and an alternative hypothesis, with the Bayes factor providing evidence in support of one hypothesis (the numerator hypothesis) relative to the other hypothesis (the denominator hypothesis). We will notate a Bayes factor in support of the null hypothesis (numerator) relative to the alternative hypothesis (denominator) as \(BF_{01}\), where 01 in the subscript indicates the null (0) over the alternative (1). We will use \(BF_{10}\) to notate a Bayes factor for the alternative over the null. A Bayes factor of greater than 1 supports the numerator hypothesis while a Bayes factor of less than 1 supports the denominator hypothesis. We will interpret Bayes factors between 1/3 and 3 as being ambiguous about which hypothesis is supported, while Bayes factors greater than 3 or less than 1/3 are taken to meaningfully support one hypothesis to the other.

On half of the trials the probe feature was a color and on the other half of the trials it was an orientation. To determine whether the probe feature type had an effect on accuracy, we performed a 2 (probe feature type) X 3 (test situation) within-participants Bayesian ANOVA (Rouder, Morey, Speckman, & Province, 2012) performed with the BayesFactor (Morey & Rouder, 2015) package for R (R Core Team, 2016). We found clear evidence against a main effect of probe feature type \((BF_{01} = 6.7)\) and against an interaction between probe feature type and test situation \((BF_{01})\).
Figure 2.3: Accuracy (left) and response latency (right) data for Experiment 1a. Error bars are 95% credible intervals. The test situations each allow for two response strategies, which are Matching (M), Deduction (D), and Familiarity (F). Test situation 1, array size 1 contains no data because a single item array cannot allow for both Matching and Deduction. The legend applies to both panels.

Given that the probe feature type had no effect, we collapsed across probe feature type for all further analyses.

Accuracy and response latency data are plotted in Figure 2.3. We examined accuracy as a function of array size and test situation with Bayesian ANOVA. We found a main effect of array size, $BF_{10} = 2.4 \times 10^9$. We also found a main effect of test situation, $BF_{10} = 740$. We found no interaction, $BF_{01} = 4.2$. This pattern of results was the same when array size 1 (1 non-hidden object in the array) was excluded from the analysis, which indicates that the pattern of results is not due to array size 1 distorting the results. We performed the same analysis for response latency that we did for accuracy. When all array sizes were included, we found both main effects (array size $BF_{10} = 10$; test situation $BF_{10} = 7.9 \times 10^5$), but no interaction, $BF_{01} = 4.2$. When array size 1 was excluded, we instead only found a main effect of test situation, $BF_{10} = 1.7 \times 10^{12}$, but no other effects.
2.4.1 Model Selection

We now turn to modeling our data. In the model for this task, $\lambda$ is a nuisance parameter that we would prefer to estimate in as parsimonious a way as possible. We fit four models that varied in terms of how flexible $\lambda$ was. Model 1 was that $\lambda = 0$ for all participants. Model 2 was that $\lambda$ was freely estimated for all participants. Model 3 was that $\lambda$ was freely estimated for all participants and was able to vary with array size. Model 4 was that $\lambda$ was the same value for all participants, but that value was able to vary with array size. We compared the four models using WAIC (e.g. Gelman, Hwang, & Vehtari, 2013), which is a parsimony-adjusted measure of model fit similar to AIC and BIC, but appropriate for Bayesian models. The WAIC values (calculated with $p_{WAIC}$) were 1329 for Model 1, 1250 for Model 2, 1233 for Model 3, and 1266 for Model 4. Lower values of WAIC indicate better parsimony-adjusted fit, so Model 3, which is the most flexible model, had the best parsimony-adjusted fit.

2.4.2 Parameter Estimates

We tested whether $\beta$ and $\lambda$ varied with array size using Bayesian hypothesis tests as implemented in the CMBBHT (Hardman, 2017) package for R. For $\beta$, there was a main effect of array size, $BF_{10} = 200$. As can be seen in Figure 2.4, $\beta$ is near 1 at array size 1, but drops substantially at higher array sizes. Follow-up tests showed that array sizes 3 and 4 did not differ from one another, but both differed from array size 1. For $\lambda$, there was a main effect of array size, $BF_{10} = 47$. From examining Figure 2.4, the effect seems to be due to reduced Familiarity at array size 4. Parameter estimates across multiple experiments, including this one, can be found in Table 5.1.

To understand how the parameters of our model vary at the individual participant level, we examined correlations between $K$, $\beta$, and $\lambda$. This is not related to the Deduction Discrepancy. The correlation coefficients and Bayes factors were calculated
Figure 2.4: Posterior means and 95% credible intervals for model parameters as a function of array size for both Experiment 1a (left side of dividing line) and Experiment 1b (right side of dashed vertical dividing line). The parameters reflect WM capacity ($K$), use of Deduction ($\beta$), use of Familiarity ($\lambda$), and guessing bias toward selecting the feature response ($g$; Experiment 1b only). The $K$ parameter was constrained by the model to have the same value at all array sizes in Experiment 1a. The x-values of the Experiment 1b parameters are jittered for visibility.

using the BayesMed package (Nuijten, Wetzels, Matzke, Dolan, & Wagenmakers, 2015) for R. There was a correlation between $K$ and $\beta$, $r = 0.54$, $BF_{10} = 27$. There was also a correlation between $\beta$ and $\lambda$, $r = 0.50$, $BF_{10} = 12$. Finally, there was a correlation between $K$ and $\lambda$, $r = 0.43$, $BF_{10} = 3.4$.

2.5 Discussion

This experiment serves as a conceptual replication of Hardman and Cowan (2016) performed for the purposes of examining Hypothesis 2: Inaccurate Models. For our
purposes, being specifically related to the Deduction Discrepancy, replication consists of the $\beta$ estimates of this experiment being approximately equivalent to the $\beta$ estimates of Hardman and Cowan. If not, that would be evidence in favor of Hypothesis 2, that modeling inaccuracies were the cause of the Deduction Discrepancy. Hardman and Cowan found a mean $\beta$ estimate of 0.58 while this experiment found mean $\beta$ estimates of 0.50 at array size 3 and 0.65 at array size 4. The credible intervals of these $\beta$ estimates contain the value 0.58, so this study replicates the result of Hardman and Cowan, although the strength of this conclusion is limited by the uncertainty in our $\beta$ estimates, as we discuss in the next paragraph.

One consideration in parameter estimation is the degree of precision with which a parameter value is known, which is related to the width of the credible interval for that parameter. As can be seen in Figure 2.4, for the $\beta$ parameters at array sizes 3 and 4, the credible intervals are wide, indicating a difficulty with precisely estimating the exact value of $\beta$. Hardman and Cowan (2016) had the same problem with the $\beta$ parameter. Thus, it appears that it is difficult to precisely estimate $\beta$. One reason for this is the fact that successful use of Deduction depends on both (a) an attempt to use Deduction, which depends on $\beta$ but also (b) having the necessary information in WM. Estimating how often participants attempt to use Deduction with $\beta$ is more difficult than estimating Deduction success overall because of the two-requirement nature of Deduction. Although the credible intervals for $\beta$ contain the mean estimate found by Hardman and Cowan, the credible intervals also contain many other values of $\beta$. Thus, we replicate the result of Hardman and Cowan, but that is somewhat unsurprising given the width of the credible intervals.

In an attempt to reduce the width of the $\beta$ credible intervals, we applied some constraints to the model for this experiment, which we discuss further in Appendix B. To summarize, it is possible to obtain fairly narrow credible intervals for $\beta$ in this experiment, but to do so requires some sacrifices in terms of model plausibility and
generality.

We also examined Hypothesis 1: Maintaining Representations, that maintaining items in WM in the feature-matching task is the cause of the Deduction Discrepancy. To do this, we included array size 1, which is effectively the reasoning-only task of Hardman and Cowan (in prep), but with a study-test delay in which maintaining a WM representation of the object in the sample array is required. In this experiment, we found that performance was very good at array size 1, with accuracy at 95% for trials on which Matching could be used (test situation 2) and 93% on trials for which Deduction could be used (test situation 3). This level of accuracy is slightly lower than that found by Hardman and Cowan in the reasoning-only task, but is still quite high. Thus, it does not appear that Hypothesis 1 can fully explain the Deduction Discrepancy.

Estimates of $\beta$ at array size 1 were approximately 1, which is reasonable given how accurate participants were in test situation 3, in which Deduction can be used. At higher array sizes, however, $\beta$ dropped substantially, to near 0.5 or 0.6. It appears that participants are very good at using Deduction as long as there is a very low or no WM load, but Deduction is used less at higher array sizes. It may be that Hypothesis 1 may hold, but only at higher array sizes.

Note that in the reasoning-only task of Hardman and Cowan (in prep), sample arrays could contain either 1 or 2 non-hidden objects. At array size 1 in this experiment, only 1 object was shown. The difference in the number of presented objects could have an effect. Thus, in the reasoning-only task that we used in Experiment 2, we only every presented 1 object in the sample array to better equate experimental conditions.

In the design of this experiment, array size 2 was excluded so that array size 1 would be quite distinct from the other array sizes we used. A design that uses array sizes 1, 2, and 3 could help to answer the question of how participants perform at
array size 1 when it is less distinct from the other array sizes. Such a design could also help to determine how \( \beta \) varies with array size at low array sizes.

Although the \( K \) estimates we obtained might appear to be low at first, these estimates agree with past research in which the binding between two features of an object must be known. For example, Cowan, Blume, and Saults (2013) found a WM capacity of approximately one two-feature object for which the binding was known. In addition, although Hardman and Cowan (2016) found higher estimates of \( K \) than were found in this experiment, those estimates were still low (around 1.5 objects) and may have been inflated due to the fact that Familiarity was not modeled in that study. The effect of modeling Familiarity is discussed further in the Cross-Experiment Parameter Comparison subsection in the General Discussion.

Finally, we also developed a simplified model for this task that is discussed further in Appendix C. The main advantage of the simplified model are that it is easier to use (we provide closed-form equations for the parameters of that model). In addition, its parameters correlate well with the parameters of the main model we used. The main disadvantage of the simple model is that it does not cleanly separate \( K \) and \( \beta \), providing only partial separation of those parameters.
Chapter 3

Experiment 1b

3.1 Introduction

This experiment serves as a complement to Experiment 1a by examining some of the same questions from a different perspective. In Experiment 1a, participants always had two memory-based response strategies available to them. This experiment is intended to limit the response strategies available to participants in different ways than Experiment 1a by requiring participants to respond in a different way. In Experiment 1a, participants had to choose between two response options, making the task a two-alternative forced-choice (2AFC) task. Experiment 1b instead uses a task that can be thought of as a type of same-different task. In particular, at test, participants will be asked whether the probe feature went with a single response option or whether it went with some feature other than the single response option. This task can also be thought of as a type of change-detection task, where the combination of the probe feature and the single response could be thought of as an object that may have been in the sample array, but if it was not, it represents a change. Other than the way in which participants respond, the method for this experiment is the same.
as for Experiment 1a. This experiment uses the same basic paradigm as Experiment 1a in which some sample array objects are hidden from participants.

An example of the method used in this experiment can be seen in the schematic of the method shown in Figure 3.1. If the participant believes that the probe feature and response option were from the same object, they select the feature response. If the participant believes that the probe feature and response option were from different objects, they can respond by clicking on the X-shaped response option, the rejection response. The rejection response is not meant as a hidden response option, it is meant as a negation to the feature response, which was explained to participants in the instructions. With the design of this experiment, there are five possible test situations with different response strategies, which are depicted in Figure 3.1.

### 3.1.1 Test Situations

In test situation 1, participants are tested with two features drawn from the same sample array object. Thus, Matching can be used to pair together the probe feature and the feature response, resulting in a correct response. In addition, Familiarity can be used if both features are familiar to the participant. If Familiarity is used, the response would be to select the feature response, which is correct in this test situation. Familiarity is a fairly poor response strategy to use in this case, however, because the two features could each have been drawn from two different objects but Familiarity cannot identify that fact.

In test situation 2, participants are tested with a probe feature drawn from one object and a feature response drawn from a different object. To answer correctly, participants are able to use Deduction. They can either detect that the probe feature went with a different feature than the feature response, or they can detect that the feature response went with a different feature than the probe feature. Thus, Deduction can be used if the participant has knowledge of either of the two relevant objects from
the sample array. As in test situation 1, Familiarity can be used to respond. In this
test situation, both the probe feature and the feature response can be familiar, but
Familiarity results in the feature response being selected incorrectly. Thus, the use
of Familiarity in this test situation results in systematically incorrect responses.

In test situation 3, participants are tested with a probe feature drawn from one of
the hidden objects from the sample array and a feature response drawn from an object
that was shown. For the feature response that was shown, participants are able to
use Deduction to detect that the feature response went with a different feature than
the probe feature. Note that participants cannot know the object from which the
probe feature was drawn as it was never shown to them. As a result, test situation 3
differs from test situation 2 in that participants are able to use Deduction only if they
know the object from which the feature response was drawn. Thus, the information
required to use Deduction in test situation 3 is more specific than the information
required in test situation 2. As a result, an attempt to use Deduction in test situation
3 is less likely to result in a correct response than an attempt to use Deduction in test
situation 2. Participants are also able to use Familiarity to detect that they saw the
feature response (the color orange in Figure 3.1) but that they did not see the probe
feature (pointing to the left). Thus, according to Familiarity, the probe feature and
feature response must not have gone together and the rejection response is therefore
selected, which is correct in this test situation.

Test situation 4 is the same as test situation 3, except that in this case the probe
feature is drawn from a shown object and the feature response is drawn from a
hidden object. The primary difference between test situations 3 and 4 has to do with
whether participants treat the probe feature and the feature response differently. For
example, if participants only attempt to apply Deduction to the feature response,
then participants would only be able to use Deduction in test situation 3, but not in
test situation 4.
Test situation 5 is a forced guessing condition. Participants are shown a probe feature and a feature response both drawn from objects that were hidden in the sample array. Given that participants have no information about whether or not these features went together, they have no choice but to guess. This test situation can be useful by providing information about a participant’s guessing bias: whether they tend to select the feature response or the rejection response when guessing.

3.1.2 Familiarity

The way in which Familiarity works in this model is similar to, but slightly different from, how Familiarity works in the model for Experiment 1a. When a participant attempts to use Familiarity in this experiment, if both the probe feature and the feature response are familiar, the participant believes that the features go together. If one feature is familiar and the other feature is unfamiliar, the participant believes that the features do not go together. If both features are unfamiliar, there is no basis on which to decide whether the features went together or not, so the participant does not make a response based on Familiarity.

In test situation 1, if the participant is familiar with both of the features, their response will be to correctly respond that the features were from the same object. In test situations 3 and 4, one of the features was hidden in the sample array, meaning that it is necessarily unfamiliar. If the other feature is familiar, the participant can correctly respond that the features were from different objects. In test situation 5, the forced guessing condition, participants are never familiar with either feature, so they do not systematically choose a response based on this lack of familiarity, they simply guess. Test situation 2 differs from the other situations in that use of Familiarity is actively harmful. Both of the features can be familiar, but those features were drawn from different objects. The Familiarity rule is that two familiar features should be treated as coming from the same object, so participants would incorrectly
respond that the features were from the same object. Although Familiarity would lead participants to systematically select the incorrect answer in test situation 2, in the absence of any other cue, it is reasonable for a participant to use Familiarity to respond. In addition, test situation 1 trials occur twice as often as test situation 2 trials (see the method for the explanation why), which could lead participants to use Familiarity as it would usually be successful when both features are familiar. Finally, to anticipate our results, the data suggests that participants appear to make errors in test situation 2 that are plausibly due to the use of Familiarity.

3.2 Method

There were 30 participants (17 female; mean age 18.7 years) drawn from the same participant pool as Experiment 1a.

The stimuli and timings for this experiment are the same as for Experiment 1a, except for the test display or any other minor changes mentioned below. In this experiment, the test display consists of the probe feature and two response options, as can be seen in Figure 4.1. The response option on the left is a feature that may or may not be from the same object as the probe feature, which we call the feature response. The response option on the right is an X inside of a box, which we call the rejection response. Participants were instructed to click on the feature response if they believed that it went with the probe feature or to click on the rejection response if they believed that the feature response did not go with the probe feature. As in Experiment 1a, an arrow was drawn pointing from the probe feature to the hovered response option. After responding, participants were given feedback in the same way as in Experiment 1a. Participants were given 25 points for a correct response and 0 points for an incorrect response. This was supposed to be equated with the points received by participants in Experiment 1a at array size 3, but there was a coding error.
in that experiment: Participants were supposed to receive 10 points plus 5 points for each item in the array, but the total array size (always 6) was used instead of the number of non-hidden items in Experiment 1a.

The array size on every trial was five total objects, three of which were shown to participants and two of which were hidden. Fewer array sizes were used in this experiment than in Experiment 1a because of the larger number of test situations that the data must be distributed across in this experiment. Participants performed 8 practice trials of a task identical to the main task except with no hidden objects in the sample array to acclimate them to the basic feature-matching task. They then performed an additional 8 practice trials of a task identical to the main task. After practice, participants were left to finish the experiment. Participants performed two equivalent blocks of 200 trials of the main task, with a short break in between blocks. Trials with different test situations were randomly intermixed within the two trials blocks.

In test situation 1, the correct response is the feature response. In test situations 2 through 4, the correct response is the rejection response. In test situation 5, there is no correct response. Thus, if equal numbers of the different test situations were used, participants would be correct to have a strong bias toward selecting the rejection response. To combat this, we included twice as many test situations 1 as any other type of test situation. As a result, test situation 1 was used on 2/6 of trials and each other test situation was used on 1/6 of trials. Thus, 2/6 of responses should be the feature response, 3/6 of responses should be the rejection response, and 1/6 of trials can be anything (guessing trials).

In test situation 5 (the forced guessing condition), neither answer is correct or incorrect. If participants were always given either positive or negative feedback, it could affect their performance in some way. To create the illusion that there is a correct response in test situation 5, we gave participants positive feedback on 2/3
Figure 3.1: Schematic of the method for Experiment 1b. The sample array contained 5 total objects, 2 of which were hidden and 3 of which were shown. At test, the participant determined whether the probe feature (the single outlined triangle) went with the feature response (colored circle), in which case the participant should click on the feature response, or not, in which case the participant should click on the square containing an X. There were five test situations that differed in terms of which response strategy could be used, indicated below the test situation (Familiarity was also possible in all test situations except for the forced guessing situation).

of trials. In part this was to help maintain the illusion that there were no forced guessing trials: If participants received ambivalent feedback when they felt as though they were being forced to guess, it could lead them to believe that they were being forced to guess. By receiving above-chance feedback, it could lead them to believe that they actually did know something and were not being forced to guess.

3.3 Model

The assumptions of the model for this experiment are the same as for the Experiment 1a model. Refer to the process model trees in Figure 3.2 in this section. Additional information on the model specification is given in Appendix A.
Experiment 1a used a 2AFC task, in which participants should not have a systematic bias toward selecting a given response option. In this experiment, however, the two responses have different identities and participants may have a bias toward selecting either the feature response or the rejection response when guessing. Thus, this model requires a guessing bias parameter that was not required in the Experiment 1a model. We call this parameter $g$ and it reflects the probability that a participant will select the feature response when guessing, which is the same as a bias toward believing that the probe feature and feature response were drawn from the same object.

In test situation 1, participants can use either Matching or Familiarity. Beginning from the node marked ”1”, participants proceed to use Matching if they have the necessary information in WM. With probability $K/N$, the object needed to use Matching is in WM and the participant selects the Same response. If Matching fails due to lack of information, the participant is able to correctly respond with Familiarity with probability $\lambda$. In this test situation, both features are familiar, so Familiarity leads participants to select the Same response, which happens to be correct. If Familiarity fails, the participant has exhausted all memory-based response strategies and must guess. When guessing, the Same response is selected with probability $g$.

In test situation 2, participants can use either Deduction or Familiarity. Beginning from the node marked ”2”, participants proceed to attempt to use Deduction with probability $\beta$. The probe feature and feature response are drawn from two different objects, so if a participant attempts to use Deduction, knowledge of either of two objects will allow Deduction to be successful. This is captured in the model with the two nodes marked ”$A_1$” and ”$A_2$” (Deduction attempted 1 and 2), from which the participant has some probability of knowing the object that will allow Deduction to succeed. From the ”$A_1$” node, the participant knows the first necessary object with probability $K/N$. If the participant lacks that knowledge, they know the other object with probability $K/(N - 1)$. The logic for the denominator being $N - 1$ is the same.
logic used for the probability of knowing the information needed for Deduction in test situation 1 of the Experiment 1a model. If Deduction fails or was not attempted, participants proceed the node marked ”F” and use Familiarity with probability $\lambda$. In this test situation, both features are familiar, so Familiarity leads participants to select the Same response, which happens to be incorrect. If Familiarity is not used, the participant guesses, selecting the Same response with probability $g$.

In test situations 3 and 4, participants are able to use Deduction or Familiarity. In both test situations, one of the features was presented to participants and the other was not, which results in the model being the same for both test situations. Beginning from the node marked ”3/4”, participants proceed to attempt to use Deduction with probability $\beta$. If they do so, they move to the node marked ”A” and have the information necessary to use Deduction in WM with probability $K/N$. If Deduction fails or was not attempted, participants proceed the node marked ”F” and use Familiarity with probability $\lambda$. In this test situation, one feature is familiar and the other is unfamiliar, so Familiarity leads participants to select the rejection response, which happens to be correct. If Familiarity fails, the participant guesses, selecting the Same response with probability $g$.

In test situation 5, participants have no information and must guess, selecting the Same response with probability $g$.

3.4 Results

We tested whether accuracy and response latency varied with test situation with Bayesian ANOVA. For the accuracy test, we excluded test situation 5 as accuracy is undefined in that test situation. We found that accuracy varied with test situation, $BF_{10} = 12$. We found ambiguous evidence about whether response latency varied
Figure 3.2: Binomial process trees for the Experiment 1b model. The number in the farthest-left node is the test situation that that tree applies to. Test situations 3 and 4 use the same tree. The round nodes labeled "M", "D", "F", and "G" indicate the start of the Matching, Deduction, Familiarity, and Guessing response strategies, respectively, with the branches following that node giving probabilities of different outcomes of that strategy. The round nodes marked "A", "A_1", and "A_2" indicate that Deduction has been attempted. A square node indicates a response, where "S" indicates a Same response (the feature response is selected) and "D" indicates a Different response (the rejection response is selected). The Same response is only correct in test situation 1.
One question of interest is whether participants use Deduction with the probe feature, feature response, or both. We can examine this question by testing whether accuracy or response latency differ between test situations 3 and 4. We found weak evidence that favored the hypothesis that test situations did not differ for accuracy, $BF_{01} = 1.7$, or response latency, $BF_{01} = 2.8$. If performance is the same in test situations 3 and 4, it would suggest that participants may use Deduction with both the probe feature and the feature response. Due to the lack of clear difference between test situations 3 and 4 in our data, we chose to collapse those test situations for the purposes of modeling the data.

In Experiment 1a at array size 3, average response latency was 1591 ms. In Experiment 1b, average response latency, excluding test situation 5, was 1425 ms. In addition, accuracy was 72% in Experiment 1a and 73% in Experiment 1b. Experiments 1a and 1b are very similar except for the way in which participants respond.
The major difference between change-detection and 2AFC tests is the number of responses that must be considered. In change-detection, only one response must really be considered, as one response is the negation of the other. In 2AFC, both responses must be considered. Whatever the reason for the differences in response latency between the tasks, the response latency and accuracy differences suggests that the change-detection task is somewhat easier than the 2AFC version of the task.

3.4.1 Model Selection

In the model for this experiment, $\lambda$ and $g$ are nuisance parameters, so we examined the potential to constrain the complexity of the model by using single $\lambda$ and $g$ parameters that were shared by all participants. Crossing whether $\lambda$ and $g$ were shared creates four model variants. We fit each of these variants and compared them using WAIC. For the model with free $\lambda$ and free $g$, WAIC = 1,032. For the model with free $\lambda$ and shared $g$, WAIC = 1,473. For the model with shared $\lambda$ and free $g$, WAIC = 1,084. For the model with shared $\lambda$ and shared $g$, WAIC = 1,523. Thus, the winning model according to WAIC is the most flexible model, with free $g$ and $\lambda$ parameters for each participant. The following analyses were all performed with the winning model.

3.4.2 Parameter Estimates

Posterior means and credible intervals of the parameters in this experiment are plotted in Experiment 2.4. Correlations between $K$, $\beta$, and $\lambda$ were examined as in Experiment 1a. We found that $\beta$ and $\lambda$ were correlated, $r = 0.62$, $BF_{10} = 220$. In addition, $\beta$ and $g$ were correlated, $r = 0.68$, $BF_{10} = 1.7 \times 10^3$. There was evidence against any other correlation being present, all $BF_{10} < 0.69$ and all $|r| < 0.31$.

To determine whether $K$, $\beta$, and $\lambda$ were the same in Experiments 1a and 1b, we compared those parameters between experiments. Mean parameter estimates
with credible intervals are plotted in Figure 2.4 and the means are given in Table 5.1. In this experiment, the only array size was 3, so we used the parameter values from array size 3 from Experiment 1a for greatest comparability. For $K$, we found ambiguous evidence in favor of a difference between experiments, $BF_{10} = 2.3$. For $\beta$, we found ambiguous evidence in favor of no difference between experiments, $BF_{01} = 2.1$. Finally, for $\lambda$, we found evidence in favor of no difference between experiments, $BF_{01} = 7.6$. Thus, although there was a suggestion, albeit ambiguous, that there was a difference between experiments in the $K$ parameter, overall there was no clear evidence of a difference between experiments.

### 3.5 Discussion

This experiment produced similar parameter estimates to Experiment 1a. Although it uses a different method and model, this experiment substantially replicates the results of Experiment 1a. Again, the credible interval for $\beta$ contains the value of 0.58 for $\beta$ found by Hardman and Cowan (2016). Again, however, note that the width of the credible interval limits the strength of this conclusion as many values of $\beta$ are within the credible interval. This result is discussed further in the general discussion.

In test situation 2, participants are able to use Deduction with either of two different objects. This suggests that accuracy in that test situation should be better than in test situations 3 and 4, as Deduction can only be used if the participant has knowledge of one specific object in those test situations. Thus, if Deduction operates as we believe and there is no Familiarity, accuracy should be better in test situation 2 than in test situations 3 and 4. We found, if anything, worse performance in test situation 2 than in test situations 3 and 4. The first explanation for the lower accuracy in test situation 2 is the fact that using Familiarity in test situation 2 is actually systematically harmful to accuracy. Lower accuracy in test situation 2
than in test situations 3 and 4 could be used as an index of the use of Familiarity in responding. A secondary explanation is that Deduction does not work as we expect in test situation 2. Perhaps, rather than attempting to use Deduction with both the probe feature or feature response as the model assumes, participants only use Deduction with one of the probe feature or the feature response on a given trial. Still, this is at best a partial explanation as the use of only one of the probe feature or feature response would hurt performance in test situations 3 and 4 as well, not just test situation 2.

It is not clear that the forced guessing condition (test situation 5) provides a particularly good index of participants’ guessing bias, $g$. The proportion of feature responses in that condition was highly variable, moreso than the other conditions, as can be seen in the middle panel of Figure 3.3. This suggests that participants might have idiosyncratic responding patterns when faced with a condition that they recognize as or suspect to be forced guessing. Thus, participants may be using a response strategy other than guessing, but the exact nature of that strategy is difficult to ascertain, in part because there is no correct answer in the forced guessing condition. Some participants may use the rejection response as we intended, as meaning ”not the feature response”, but also to mean ”I don’t know the response”, which could increase the variability. Alternately, participants really do have highly variable guessing biases ($g$), but the variability of guessing biases is less apparent in the other test situation in which guessing accounts for a smaller proportion of responses.

There is one pattern in the data that may appear to contradict the model estimates of $\beta$. This pattern is that accuracy in test situation 1 was nearly equal to accuracy in test situations 3 and 4. Based on the model trees in Figure 3.2, if $\beta = 1$, then test situation 1 and test situations 3 and 4 are nearly identical. As a result, equal accuracy between test situations 1, 3, and 4 indicates that $\beta$ must be approximately 1. There is one important difference, however, between test situation 1 and test situations 3
and 4, which is that the correct response differs between those test situations and is affected by the guessing bias. The guessing bias was toward selecting the rejection response, which is correct in test situations 3 and 4 but incorrect in test situation 1. Thus, the guessing bias increases the accuracy in test situations 3 and 4 while decreasing the accuracy in test situation 1. As a result, equal accuracy between test situations 1, 3, and 4 does not imply that $\beta$ must be 1.
Chapter 4

Experiment 2

4.1 Introduction

In this experiment, we investigated Hypothesis 3: WM Load, that the Deduction Discrepancy is due to a competition for WM resources that manifests when WM is loaded. In the reasoning-only task, WM is lightly loaded, but in the feature-matching task at moderate to high array sizes, WM is heavily loaded. Hardman and Cowan (2016) and our Experiments 1a and 1b found that the amount of Deduction used by participants was moderate. This may be because, at test, participants split their resources between maintaining WM representations of items from the sample array and performing Deduction. In the reasoning-only task, maintaining the WM load is not required because the stimuli are always visible to participants, so they may be able to devote more resources to using Deduction.

To examine how WM load affects reasoning, we used a dual-task design that combines the feature-matching and reasoning-only tasks. On each trial of the dual task, participants were first shown the sample array for the feature-matching task. The sample array was then replaced by the reasoning-only task display and participants
made their response to the reasoning-only task. Finally, the feature-matching task was completed, with participants being tested on their memory for the original sample array. This trial structure is summarized in Figure 4.1. We also used standard single-task versions of the feature-matching and reasoning-only tasks to establish baseline performance on those tasks for this sample of participants.

We were able to measure how effectively participants used Deduction by examining performance in the reasoning-only task. We investigated the effect of WM load on the reasoning-only task by manipulating how many objects participants were required to remember in the feature-matching task. The reasoning-only single task establishes the level of performance when the participant is under no WM load. If WM load affects reasoning, we would expect reduced reasoning-only task performance when participants are placed under WM load. More importantly, Hypothesis 3 predicts that Deduction and Matching respond differently to WM load. Thus, Hypothesis 3 predicts an interaction between WM load and the type of response strategy required in the reasoning-only task. For example, we might expect that WM load hurts Deduction more than Matching. Such an interaction would support Hypothesis 3, but a failure to find an interaction or to find a different form of interaction would reject Hypothesis 3.

The design as discussed allows for the assessment of the effect of WM load on reasoning ability, which is of primary interest. Given this design, however, it is straightforward to examine a related issue, which is the effect of reasoning on WM performance. If Deduction uses WM resources, then it is reasonable to think that when a participant uses Deduction in the reasoning-only task, it might displace memory items that were being maintained for the feature-matching task. This would result in reduced performance in the feature-matching task when Deduction is required in the reasoning-only task versus when Matching is required. This logic holds as long as Deduction requires more WM resources than Matching. We investigated this issue by
examining how the use of Deduction versus Matching in the dual-task reasoning-only task affected performance on the feature-matching task. In addition, the feature-matching single task establishes baseline performance when no reasoning is required during maintenance.

4.1.1 Related Work

In related work, Baddeley and Hitch (1974) used tasks in which participants were required to do processing/reasoning under WM load. In one example of such a task, participants judged the truth value of sentences while either (a) remembering a random 6-digit number or (b) repeating the numbers 1 through 6 to themselves. When the six digits to be repeated are random, it requires WM to maintain that information, but when the six digits are a simple sequence, it only takes some attentional resources. Baddeley and Hitch measured how long it took participants to judge the truth value of the sentences and found that participants were slower when WM was loaded. Thus, it appears that a concurrent WM load affects reasoning about unrelated stimuli.

Many researchers have done studies similar to Baddeley and Hitch (1974) in that the tasks combine storage of information in WM while performing a secondary processing task (e.g. Barrouillet, Bernardin, & Camos, 2004; Unsworth & Engle, 2004; Vergauwe, Camos, & Barrouillet, 2014). For example, in the operation span task (e.g. Unsworth & Engle), which is a type of complex span task, participants remember words while performing a secondary task in which they do mathematical operations. The memory and processing tasks are clearly unrelated. This is typical of complex span tasks. An important factor that distinguishes our Experiments 1a and 1b from past research is that in most past research the processing task is unrelated to the storage task.

An example of a complex span task in which the memory and processing tasks are
related comes from Vergauwe et al. (2016). In that study, participants remembered letters for eventual recall at the end of the trial, while the intervening processing task was to decide whether a probed letter was in WM. Although Vergauwe et al. used a processing task related to the contents of WM, there was no component of Deduction in that study. Thus, one of the important contributions of the feature-matching task is that it examines specifically Deduction performed using the contents of WM in one task that combines both reasoning and memory. This combined use of WM and Deduction should be fairly closely related to real-world reasoning, in which the premises and partial results of reasoning must be maintained in WM.

4.1.2 Mental Models

One possible reason why concurrent WM load might affect Deduction is the possibility that the process of Deduction requires WM storage resources. If this is the case and WM is full, it might be difficult or impossible to use Deduction because of the lack of available WM space. Alternately, the use of Deduction could force other items out of WM. Thus, participants may not use Deduction if WM is full. To allow us to draw on research related to propositional reasoning, we will show that the feature-matching task test situation can be conceptualized as a kind of propositional argument (e.g. Prado, Chadha, & Booth, 2011). Imagine that the participant remembers green/up. If the probe feature is red and the response options are left and up, the following argument can be made: 1. Green and up were paired together. 2. All colors and shapes have unique pairings (i.e. the pairings are mutually exclusive). 3. From 1 and 2, the orientation up was green so up cannot have been red. 4. One of the response options is always correct. 5. From 3 and 4, the only response option other than up is the orientation left, so left must have been red.

In practice, the steps in this argument are simplified somewhat by the fact that propositions 2 and 4 are always the same regardless of what the sample array or test
display contain. As argued by Oberauer (2010), some kinds of information related to performing tasks are held in a separate storage system, a "task set", than the stimuli that are to be processed. This suggests that propositions 2 and 4 might be stored in the task set separately from the memoranda that are stored in the traditional WM capacity. In that case, propositions 2 and 4 would not compete for WM capacity with the stimuli.

Proposition 1 naturally takes space in WM because it is the memory of the original object. Proposition 3 may also require space in WM, as suggested by research on mental models used for deductive reasoning (Johnson-Laird, Byrne, & Schaeken, 1992). Barrouillet and Lecas (1999) argue that some types of reasoning require the creation in WM of mental models of possible hypothetical states of the world. If this is the case, then the test situation in our tasks could require WM resources for both storing the necessary information from the sample array and also representing hypothetical states of the world. This additional burden could result in performance on the reasoning-only task being lower when a secondary WM load is imposed.

4.2 Method

In this experiment, 32 participants (17 female; mean age 19.5 years) from the same participant pool as the other experiments took part. In this experiment, some participants had near-chance accuracy on some of the tasks. We excluded participants who had less than 55% accuracy on one task (the dual task or the feature-matching or reasoning-only single tasks). This resulted in two participants being excluded, leaving 30 participants for analysis.

As an overview of the method, participants performed three tasks in this experiment: (a) The dual task, with the reasoning-only task embedded within the feature-matching task, (b) the reasoning-only single task, and (c) the feature-matching single
Figure 4.1: Schematic for the Experiment 2 tasks. Top: Reasoning-only task. Bottom: Feature-matching task. The dual task is created by inserting the reasoning-only task into the maintenance interval in the feature-matching task as indicated by the dashed arrows. In the feature-matching single task, the maintenance interval is 5 s long; it is broken into three segments to show where the embedded reasoning-only task is inserted in the dual task. The reasoning-only task example shown here is a trial on which Deduction is required.

Task. The design of these tasks is summarized in Figure 4.1, in which both tasks are shown individually and the dual task is indicated. The specifics of each task will be described further. At the beginning of the experimental session, participants practiced all three tasks in a fixed order: feature-matching single task, reasoning-only single task, and the dual task. Participants performed 8 practice trials of each task, with instructions before each set of practice trials. Following practice, participants performed one trial block of each of these three tasks in a random order. Participants performed 50 trials of the feature-matching single task, 100 trials of the reasoning-only single task, and 100 trials of the dual task.
4.2.1 Feature-Matching Single Task

The feature-matching single task was very similar to the tasks used in Experiments 1a and 1b, with the major exception that the feature-matching task used in this experiment did not involve hiding any of the objects in the sample array. This makes this task very similar to the feature-matching tasks used by Hardman and Cowan (2016) and Hardman and Cowan (in prep). This task is shown at the bottom of Figure 4.1. Participants were shown a sample array containing either 2 or 3 objects for 1 s. Then the screen was blanked for the 5 s maintenance interval. Following maintenance, the test display with one probe feature and two response options was shown to participants until they made their response. Once a response was made, feedback was shown. Feedback was the same as in Experiments 1a and 1b. A correct response was worth 20 points at array size 2 and 25 points at array size 3. An incorrect response was worth 0 points. The feedback display remained on screen until the participant initiated the next trial by clicking anywhere on screen. When the next trial was initiated, the screen was blanked for 500 ms before the sample array was presented.

4.2.2 Reasoning-Only Single Task

The reasoning-only single task is shown at the top of Figure 4.1. The reasoning-only task is essentially the same as the feature-matching task, except in two important ways. The first was that the sample array and test display are shown on screen at the same time. In this design, memory load and forgetting are non-issues because the relevant information is always available to participants. The sample array was presented on the left side of the screen with the text ”You know:” presented above it. The test display, which was identical to the test displays used in the feature-matching task, was presented on the right side of the screen. Both the sample array and test
display were surrounded by the outlines of grey rectangles, which helped to visually isolate the sample array and test display.

The second way in which the reasoning-only task differed from the feature-matching task was that some of the objects in the sample arrays were hidden, just as in Experiments 1a and 1b. We explained to participants that this was to simulate a case in which some of the objects from the original sample array were forgotten by the participant. As was discussed in the Reasoning-Only Task subsection in the Introduction, hiding some of the objects creates test situations in which the participant is forced to use either Matching or Deduction.

In order to create the dual task, we wanted to make the reasoning-only task fit into the fixed-length maintenance interval in the feature-matching task. To accomplish this, we required participants to make a speeded response in the reasoning-only task. At the start of a reasoning-only trial, the sample array and test display were simultaneously presented. From the time of this presentation, participants were given 3.5 s to make their response. This time limit was somewhat higher than average response times for adults found by Hardman and Cowan (in prep), who did not use speeded response times. We chose 3.5 s so that participants would not be under strong pressure to respond quickly, because strong time pressure might cause participants to not use Deduction. As soon as a response was made, feedback was presented. Feedback was the same as in the feature-matching task. A correct response was worth 20 points if Matching was required or 25 points if Deduction was required. This small difference in the value of the types of trials was used to encourage participants to attempt to use Deduction. An incorrect response was worth 0 points. Once a response was made and feedback was presented, participants were forced to wait the remainder of the 3.5 s before they were allowed to advance to the next trial. Once that time period had elapsed, they could initiate the next trial by clicking anywhere on screen. Once the next trial was initiated, there was a 500 ms blank before the sample array and test
display were presented. If a participant failed to make a response during the 3.5 s response window, they were given feedback indicating an incorrect response.

4.2.3 Dual-task: feature-matching task with embedded reasoning-only task

In the dual task, participants were shown the feature-matching sample array for 1 s, followed by a 1 s blank. Then the reasoning-only task was presented and participants were given 3.5 s to respond, just as in the single task. Once participants made their reasoning-only response, feedback for that task was presented and remained on screen for the remainder of the 3.5 s. Once the full 3.5 s for the reasoning-only task elapsed, there was a 500 ms blank and then the feature-matching test display was shown. Participants made an unspeeded response in the feature-matching task and received feedback on their response. Once feedback had been presented, participants initiated the next trial by clicking anywhere on screen. The dual task design can be seen in Figure 4.1 by inserting the reasoning-only task into the 3.5 s window in the feature-matching task maintenance interval.

Note that the timings of the dual task match those of the single tasks. In particular, the feature-matching single task had a 5 s maintenance interval. That same amount of time was filled in the dual task with a 1 s blank, 3.5 s reasoning-only task, and 500 ms blank. Thus, maintenance duration was equated between feature-matching single and dual tasks.

The stimuli used for the dual task were the same as for the single tasks. In the dual task, the colors and orientations on a given trial never occurred in both the feature-matching and reasoning-only tasks, which was done to prevent information from one task affecting performance on the other task. Participants were instructed about this and told to treat the two tasks as being totally separate.

There is a minor difference between the single and dual reasoning-only tasks with
respect to feedback. In the dual task, once the 3.5 s had elapsed, the display was automatically blanked before continuing to the feature-matching test. In the reasoning-only single task, the feedback display remained on screen until participants chose to initiate the next trial. Thus, in the single task, participants have more time to view the feedback display. Another difference is that the reasoning-only single task was initiated by participants while the dual-task variant started automatically after the post-sample-array blank. We believe these differences are negligible when compared to the main differences between the single- and dual-task conditions.

4.3 Results

Responses in the reasoning-only task were speeded. As a result, participants sometimes failed to respond within the allowed response window. In the reasoning-only single task, there was no response on 0.87% of trials. In the reasoning-only dual task (embedded within the feature-matching task), there was no response on 1.03% of trials. Given that these proportions of non-response were comparable, we did not consider non-responding to be different between the single- and dual-task reasoning-only tasks. No participants had excessive rates of non-response, with the highest non-response rate being 5% averaged across reasoning-only variants. When no response was made, that may have happened because the participant was ignoring the task. In the dual task, this is problematic because participants might be able to improve feature-matching performance by ignoring the reasoning-only task. Thus, we excluded all trials for which no response was made on the reasoning-only task from the analyses. We did this for both the single- and dual-task reasoning-only tasks and, in addition, dual-task feature-matching trials on which no response was made in the embedded reasoning-only task.
4.3.1 The Effect of Memory on Reasoning

To examine how memory load affects reasoning, we analyzed reasoning-only task performance as a function of the memory load of the feature-matching task. In the single-task condition, the reasoning-only task was not embedded within a feature-matching task, which results in a WM load of 0. Thus, we will treat WM load as a factor with three levels: 0 WM load in the reasoning-only single task and WM loads of 2 and 3 in the dual-task. If WM load affects reasoning, we would expect a main effect of WM load on reasoning-only task performance.

We also want to know whether Matching and Deduction require differing amounts of WM resources. If Matching and Deduction require differing amounts of WM resources, we would expect that there would be a differential effect of memory load on the two response strategies. To examine this, we would look for an interaction between memory load and the response strategy required by the reasoning-only task.

Accuracy and response latency data from the reasoning-only task are shown in the bottom two panels of Figure 4.2. We performed a 2 (response strategy) X 3 (memory load) Bayesian ANOVA on reasoning-only accuracy. We found evidence for a main effect of memory load, BF\textsubscript{10} = 7.4, which indicates that the reasoning-only task is hurt by a secondary WM load. We also found a main effect of response strategy, BF\textsubscript{10} = 2.4 \times 10^5, with participants having worse accuracy when Deduction was required. There was ambiguous evidence against an interaction between memory load and response strategy, BF\textsubscript{10} = 0.36, which indicates that Matching and Deduction do not have differing WM requirements. For response latency, we found the same pattern: two main effects, but no interaction (memory load BF\textsubscript{10} = 3.1 \times 10^8, response strategy BF\textsubscript{10} = 4.6 \times 10^{27}, interaction BF\textsubscript{10} = 0.11).

To clarify which memory loads differed from one another, we performed follow-up tests in which we did pairwise comparisons between memory loads 0 and 2 and memory loads 2 and 3. There was no interaction between memory load and response
strategy, so we collapsed across the response strategy for this test. We found a
difference between memory loads 0 and 2, BF_{10} = 6.0, but ambiguous evidence for
a difference between memory loads 2 and 3, BF_{10} = 1.7. Thus, it appears that
some memory load affects reasoning-only accuracy, but the magnitude of that load
is irrelevant. There are some possible limitations of our WM load manipulation that
we address in the Discussion of this experiment.

When we examined response latency in the same way as accuracy, we found a
slightly different pattern of results. The statistical conclusions match what is imme-
diately apparent from a visual examination of the bottom-right panel of Figure 4.2,
namely that there are main effects of memory load and response strategy, but no
interaction between the two (all Bayes factors greater than 3 in favor of the conclu-
sion). As can be seen in the figure, a higher memory load in the feature-matching
task increases response latencies in the reasoning-only task. In addition, for trials
on which the Deduction response strategy was required, participants responded more
slowly than when the Matching response strategy was required.

4.3.2 The Effect of Reasoning on Memory

To examine how reasoning affects memory, we need to examine how feature-matching
task performance changes as a function of (a) single-task vs dual-task condition and
(b) within the dual-task condition whether the reasoning-only task required Matching
vs Deduction. A secondary question is whether memory load, operationally defined
as feature-matching task array size, interacts with task condition. Feature-matching
task accuracy and response latency are plotted in the top two panels of Figure 4.2.

We performed a 2 (array size) X 3 (task condition) Bayesian ANOVA with feature-
matching accuracy as the dependent variable. We found evidence for a main effect of
task condition, BF_{10} = 2.1 \times 10^{18} and a main effect of array size, BF_{10} = 3.7 \times 10^{11}.
There was also evidence for an interaction between task condition and array size, BF_{10}
Figure 4.2: Experiment 2 accuracy and response latency for the feature-matching task (top) and reasoning-only task (bottom). For the feature-matching task, the x-axis indicates what type of reasoning was required during the maintenance interval, which was none in the single task and either Matching or Deduction in the dual task, while connect lines indicate the array size. For the reasoning-only task, the x-axis indicates secondary memory load, which was 0 in the single task and either 2 or 3 in the dual task, while connect lines indicate what kind of reasoning was required in the reasoning-only task.
= 4.8. For response latency, main effects of both task condition, \( BF_{10} = 6.7 \times 10^3 \), and array size, \( BF_{10} = 9.2 \times 10^9 \), were present, but the interaction was not, \( BF_{10} = 1.0 \). Interestingly, the main effect of task condition is due to the fact that participants responded more slowly in the single task condition.

To determine whether there was an overall effect of single- vs dual-task condition, we examined how feature-matching accuracy and response latency varied with task condition. We performed a 2 (array size) X 2 (task condition: single or dual) ANOVA separately for each of accuracy and response latency. For accuracy, we found a main effect of task condition, \( BF_{10} = 1.3 \times 10^{16} \), a main effect of array size, \( BF_{10} = 1.6 \times 10^7 \), and an interaction, \( BF_{10} = 7.0 \). Given the type of interaction, both main effects are interpretable. For response latency, we found a main effect of task condition, \( BF_{10} = 4.9 \times 10^2 \), a main effect of array size, \( BF_{10} = 1.7 \times 10^7 \), but ambiguous evidence in favor of an interaction, \( BF_{10} = 1.4 \). The most important results here are that there are main effects of single- vs dual-task condition for both accuracy and response latency.

To examine the difference in feature-matching performance between dual-task situations in which the reasoning-only task required Matching vs Deduction, we performed an ANOVA in which we did not include the single-task condition. As a result, we performed a 2 (array size) X 2 (reasoning type: Matching or Deduction) ANOVA separately for each of feature-matching accuracy and response latency. For accuracy, we found evidence for a main effect of array size, \( BF_{10} = 2.8 \times 10^{10} \). We found only ambiguous evidence in favor of a main effect of reasoning type, \( BF_{10} = 1.7 \), and ambiguous evidence against an interaction, \( BF_{10} = 0.41 \). The same pattern repeated for response latency, for which we found a main effect of array size, \( BF_{10} = 7.4 \times 10^6 \), but no main effect of task condition, \( BF_{10} = 0.48 \), and no interaction, \( BF_{10} = 0.29 \). Thus, there does not appear to be a difference in feature-matching performance depending on which type of reasoning was required in the embedded reasoning-only task.

To examine the effect of accuracy in the reasoning-only task, we repeated these
analyses only for trials on which the reasoning-only response was correct. Since accuracy in the reasoning-only task was quite high, this only resulted in a small proportion of trials being removed for this analysis. We found the same pattern of results as with the full data set.

4.4 Discussion

This experiment examined Hypothesis 3: WM Load. If WM load negatively impacts Deduction use, we would expect that putting participants under a WM load while they performed the reasoning-only task would hurt Deduction use in that task. We found that Deduction was hurt by having a secondary WM load, but the magnitude of the effect is too small to account for the Deduction Discrepancy, which we expand on in the next paragraph. In addition, Hypothesis 3 was specific to Deduction being affected by WM load, but we found that both Matching and Deduction in the reasoning-only task were equally hurt by a WM load. Thus, Hypothesis 3 is not supported by this experiment.

In psychometric modeling experiments (e.g. Experiments 1a and 1b), we have consistently found $\beta$ estimates near 0.5. If the this moderate value is largely due to WM load at test, we would expect that in the reasoning-only dual task that participants would be able to use Deduction about half of the time. Assume that WM load causes Deduction to be used only 50% of the time it could be used. If that is true, we would then expect that in the dual-task reasoning-only task, participants would use Deduction on only half of trials. Participants respond correctly when Deduction is used, which is half of trials, and guess correctly on half of the remaining trials, which results in 75% accuracy on the dual-task reasoning-only task when Deduction is required. Instead, we found accuracy above 90% on trials for which Deduction was required. Thus, although we found in this experiment that WM load affects performance on Deduction trials of the reasoning-only task, the magnitude of the effect is too small
to be consistent with the moderate $\beta$ estimates from other experiments. Although WM load may be a partial explanation for the Deduction Discrepancy, an additional explanation above and beyond WM load is required to explain the moderate amounts of Deduction used in the feature-matching task.

In the feature-matching task, at test, participants have a WM load that is directly useful to them. In this experiment, with reasoning-only embedded into a feature-matching trial, the objects comprising the WM load from the feature-matching task are irrelevant to the reasoning-only task. Given that a relevant WM load could be used for the purposes of Deduction, we would expect that an irrelevant WM load would hurt Deduction at least as much as a relevant WM load of equal size, if not more. We found that Deduction in the reasoning-only task was only slightly hurt by having an irrelevant WM load. Thus, the amount that Deduction is hurt by an irrelevant WM load establishes a ceiling on how much a relevant WM load should hurt Deduction. Thus, this experiment is a good test of whether a WM load impacts Deduction and thus a good test of Hypothesis 3.

If Deduction requires WM resources (Barrouillet & Lecas, 1999), rather than a WM load hurting Deduction in the reasoning-only task, perhaps using Deduction in the reasoning-only task hurts the WM representations of items that are being maintained for the feature-matching task. Thus, perhaps accuracy in the feature-matching task should depend on the reasoning performed in the embedded reasoning-only task. We found that feature-matching accuracy depended heavily on single vs dual-task condition, as can be seen in Figure 4.2. We did not, however, find any difference in feature-matching accuracy depending on whether participants had to use Matching vs Deduction in the embedded reasoning-only task. One possible explanation for this result is that both Matching and Deduction require WM resources to perform. Imagine that a participant remembers green/up. If the probe feature is green and the response options are left and up, the following argument can be made: 1. Green
and up were paired together. 2. All colors and shapes have unique pairings. 3. The only possible correct response option is up. Steps 1 and 2 here are the same as steps 1 and 2 for performing Deduction. The difficult steps in performing Deduction were steps 3 and 5, which Matching avoids completely. When contrasted with the steps required to perform Deduction that we listed in the Introduction to this experiment, there are fewer steps to performing Matching and the steps are easier. Thus, it seems unlikely that Matching and Deduction require equivalent WM resources to perform. It is possible, however, that much of the WM resources required for performing either Matching or Deduction are related to creating an internal representation of the sample array and test display in the reasoning-only task. If that is the case, then Matching and Deduction could be quite similar in terms of total WM resource requirements. If Deduction requires only slightly more WM resources than Matching, that could explain why there is a only weak trend suggesting that feature-matching accuracy is worse when the embedded reasoning-only task requires Deduction vs when it requires Matching.

The efficacy of our manipulation of WM load by varying the array size of the dual-task feature matching task is unclear. Based on past research, we have found that, for the color/orientation stimuli we use, nearly all participants have reached their WM capacity at an array size of two objects (Hardman & Cowan, 2016; see also Cowan et al., 2013 for more on WM capacities for two-feature objects). In addition, in Experiments 1a and 1b, we found $K$ estimates of only a little above 1 object. If WM is fully loaded at array size two, larger array sizes may not further load WM and there is only a small difference between array sizes 2 and 3. In that case our operational definition of memory load as the number of items in the sample array is not very useful. Perhaps only the difference between no load (reasoning-only single task) and some load (dual-task) is meaningful. To provide a more gradated WM load, a design with array sizes 1 and 2 could be used, as can be done with the task in
Experiment 1a. Alternately, different stimuli or tasks altogether could be used. For example, a change-detection task using letters could allow a researcher to vary WM load across a range of levels.

There appeared to be only a small effect of WM load on reasoning, with accuracy in the reasoning-only task dropping only slightly when there was a WM load. There was, however, a large effect of the embedded reasoning-only on feature-matching task accuracy. It is possible that when participants are performing the dual task, the embedded reasoning-only task was given priority over the feature-matching task, perhaps due in part to the points awarded to participants and/or relative task difficulty. Participants were given approximately equal numbers of points for correct responses in the feature-matching and reasoning-only tasks, but the reasoning-only task was easier than the feature-matching task. This combination of incentives could have led participants to strongly prioritize the embedded reasoning-only task to the detriment of the feature-matching task. Thus, the finding that there was only a small effect of WM load on reasoning could be due in large part to how participants prioritized the tasks. A different prioritization could be attained by changing how points are awarded on the tasks or by other means.

We found that performance in the feature-matching task was worse in the dual-task condition vs the single-task condition. Within the dual-task conditions, we found no difference between trials on which the reasoning-only task required Matching vs Deduction. Thus, the requirement to perform the reasoning-only task affected performance on the feature-matching task, but the specific kind of reasoning required of participants was unimportant. The pattern in the means, however, suggests that participants performed the feature-matching task slightly better when Matching was required in the reasoning-only task than when Deduction was required. A more powerful study may be able to find statistical evidence that the response strategy required by the reasoning-only task affects feature-matching performance.
Chapter 5

General Discussion

5.1 Hypothesis 1: Maintaining Representations

Our first hypothesis about the cause of the Deduction Discrepancy was that the critical difference between the feature-matching and reasoning-only tasks was the requirement that objects be maintained in WM in the feature-matching task, but not in the reasoning-only task. To examine this possibility, we included a condition in Experiment 1a that was equivalent to the reasoning-only task but with a delay between study and test. We used only a single object in the sample array, which all but guarantees that that object will be encoded into WM and maintained to the best ability of the participants. If Hypothesis 1 is correct, we would expect to find that participants would use Deduction about half of the time, resulting in about 75% accuracy for trials on which Deduction was required. Instead, we found 93% accuracy for those trials. Thus, Hypothesis 1 cannot be the sole explanation for the Deduction Discrepancy. Accuracy in the delayed reasoning-only task from array size 1 in Experiment 1a was, however, slightly lower than accuracy in the reasoning-only single task in Experiment 2, which was 96% when Deduction was required. This small
accuracy difference suggests that Hypothesis 1 may be a partial explanation for the Deduction Discrepancy.

In addition, there remains the possibility that Hypothesis 1 does not apply when the array size is very small. The $\beta$ estimates from Experiment 1a were nearly 1 at array size 1, but near to 0.5 at higher array size, which suggests that the Deduction Discrepancy may only be present at higher array sizes. It may be that for single item arrays, maintenance of items does not affect Deduction use, but that at higher array sizes maintenance does affect Deduction. Thus, Hypothesis 1 may only apply at higher array sizes. We have not attempted to use higher array sizes in the reasoning-only task, so we have no way to test this hypothesis without obtaining new data. It could be that in reasoning-only tasks that use higher array sizes Matching would still be at near ceiling accuracy but Deduction would suffer, suggesting some kind of difficulty with selecting which information to use in order to perform Deduction.

5.2 Hypothesis 2: Inaccurate Models

Our second hypothesis was that the modeling results of Hardman and Cowan (2016) were inaccurate. The modeling estimates of $\beta$ from Experiments 1a and 1b, however, agree with those of Hardman and Cowan. The finding has consistently been that $\beta$ has been near 0.5, indicating that participants attempt to use Deduction about half of they are able to. The tasks and models from Experiments 1a and 1b were more carefully designed than those of Hardman and Cowan for the purposes of accurately estimating $\beta$. Some of the improvements were controlling which response strategies could be used by participants and accounting for the use of Familiarity when responding. The consistency with which we have found moderate use of Deduction with psychometric models suggests that it may be the true result. We certainly
have no evidence in support of Hypothesis 2, that previous modeling results were inaccurate.

Note, however, that these tasks and models are not perfect and that further developments in this area would still be of value. One of the remaining issues is that the credible intervals of the $\beta$ parameters are fairly wide. For example, as can be seen in Figure 2.4, the credible interval for mean $\beta$ at array size 3 goes from approximately 0.2 to 0.8. This is problematic because (a) we might find that the mean $\beta$ is moderate, at a value of 0.51, but the uncertainty is so high that true $\beta$ could be anywhere from fairly low to fairly high and (b) as uncertainty about $\beta$ increases, the posterior distribution of $\beta$ approximately approaches some kind of symmetric distribution in the interval [0,1]. The mean of a symmetric distribution from 0 to 1 is 0.5. Thus, extreme uncertainty in $\beta$ estimates is likely to produce moderate mean $\beta$ estimates. To be confident that $\beta$ is truly moderate, it would be necessary to find a way to increase the certainty in $\beta$ estimates. We used some model-parameterization approaches in Appendix B to reduce the width of the $\beta$ credible intervals, but that approach has its limitations. In summary, although the moderate $\beta$ estimates provide evidence against Hypothesis 2, that evidence is not particularly strong due to the wide credible intervals.

5.3 Hypothesis 3: WM Load

Our third hypothesis was that Deduction use depends on the extent to which WM is loaded. This hypothesis would account for the difference between the modeling results of Hardman and Cowan (2016) and the reasoning-only results of Hardman and Cowan (in prep). In Experiment 2, we found that accuracy on the reasoning-only task was lower when there was a secondary WM load vs no WM load. This decrease, however, was the same regardless of whether Matching or Deduction was required,
which indicates that the WM load effect is not specific to Deduction. Critically, the WM effect on Deduction was small: Accuracy dropped only 1% from WM load 0 to WM load 2 and 2% from WM load 2 to WM load 3. Such a small effect of WM load cannot account for accuracy being near ceiling on Deduction trials of the reasoning-only task but $\beta$ estimates being between 0.5 and 0.6 in the feature-matching task. The WM load effect would have to be much larger to account for the Deduction Discrepancy. Thus, we have evidence against Hypothesis 3 being a complete explanation for the Deduction Discrepancy. Like Hypothesis 1, however, Hypothesis 3 may be a partial explanation for the Deduction Discrepancy.

In sum, we have evidence against all three of our hypotheses about why the feature-matching and reasoning-only tasks seem to involve very different amounts of Deduction use. That said, Hypotheses 1 and 3 may be partial explanations for the Deduction Discrepancy. In addition, Hypothesis 2 has only weak evidence against it.

5.4 New Hypotheses

Given what we have learned with this study, we have a number of new hypotheses that are potential explanations for the Deduction Discrepancy.

5.4.1 Inaccurate Metamemory

In Experiment 1a, we found that participants are able to effectively use Deduction at small array sizes, but fail to do so at higher array sizes. In particular, $\beta$ was near 1 at array size 1 in Experiment 1a, but near 0.5 at higher array sizes. It may be that at higher array sizes, participants believe that they have more information in WM than they really do. Cowan et al. (2016) used a WM task in which participants were sometimes asked to indicate how many items were held in WM and found an indication that participants believed that they had more items in WM than they
actually did. If participants believe that they have many or most of the items in
the sample array in WM, they might be less motivated to use Deduction if they feel
confident that they would be able to respond correctly using Matching, which is faster
than Deduction.

5.4.2 Degradation of Representations

Another possibility is that participants tend to attempt Deduction only after other
response strategies have been attempted. In that case, WM representations might
degrad or be lost before participants are able to use Deduction (e.g. due to decay;
Ricker & Cowan, 2010; Ricker, Spiegel, & Cowan, 2014). For example, if Matching
is attempted before Deduction, while participants are using Matching, their WM
representations could degrade during Matching so that when they attempt to use
Deduction, less information would be available in WM than when they used Matching.
In our models, we assume that whatever information is available to Matching is also
available to Deduction, but this might not be true if WM representations degrade
before Deduction is attempted. In the model for Experiment 1a, the probability that
a participant attempts to use Deduction and succeeds due to having the necessary
information in WM is \( \beta \times \frac{K}{(N - 1)} \) in test situation 1. If the number of items
available to Deduction is less than the number of items available to Matching, the
\( \frac{K}{(N - 1)} \) term will be larger than it should be because \( K \) is based in part on how
effectively participants are able to use Matching. If \( \frac{K}{(N - 1)} \) is larger than it should
be, then \( \beta \) estimates would be biased low in response.

5.4.3 Assumption of Constant \( K \)

A related issue is the possibility \( K \) may not be constant with respect to array size, as
our models assume. Some research suggests such a possibility, at least for participants
with lower WM capacities (Cusack, Lehmann, Veldsman, & Mitchell, 2009). If \( K \) is lower at higher array sizes, but the model assumes that it is constant, lower accuracy at higher array sizes may result in \( \beta \) estimates being biased low in order to account for the lower accuracy. Given the design of the task and model for Experiment 1a, however, in which \( K \) and \( \beta \) have different functions in different test situations, it should not be easy for an error in the estimation of \( K \) at higher array sizes to be pushed into \( \beta \).

### 5.4.4 Partial Objects in WM

In the delayed reasoning-only condition at array size 1 in Experiment 1a, we found that participants used Deduction well. At higher array sizes, however, participants seemed to use Deduction only moderately. This suggests that a higher WM load limits the use of Deduction. Contrary to this idea, we found that the WM capacity of participants, as estimated with \( K \), was only slightly above one object. Thus, participants’ WM was nearly full at array size 1, so it is not clear how array size 3 would result in a substantially different WM load than array size 1. One explanation of this issue is that participants might try to remember more than one object at higher array sizes, but are unable to do so in an effective way. For example, perhaps one feature of a stored object is lost, making that object useless for the purposes of Deduction, but WM is still loaded with the partial object. The partial object is still useful for Familiarity, so it makes sense for the participant to try to maintain it. This additional WM load could make Deduction more difficult, perhaps because additional WM resources are needed to hold propositions related to Deduction (Barrouillet & Lecas, 1999; Johnson-Laird et al., 1992), although we lack evidence in favor of that hypothesis (see Experiment 2). It may be possible to investigate this issue by using an instructional manipulation in which some participants are instructed to only attempt
to encode 1 or 2 objects from the sample array, which could reduce the instance of partial objects loading WM.

5.4.5 Motivation

It is possible that participants are not sufficiently motivated to use Deduction, given that there are time and, presumably, effort costs to doing so. In Experiment 2’s reasoning-only task, participants took 360 ms longer to respond when Deduction was required vs when Matching was required. In the Experiment 1a task at array size 1, the same response latency difference was 186 ms. Future research could examine the effects of motivation by either manipulating or measuring motivation. In many studies, participants are not concretely rewarded based on performance. In this study, we rewarded participants with points, but those points had no real-world value. It may be possible to motivate participants with tangible rewards, such as money, for good performance. One aspect of motivation that is relevant given the additional time demands of Deduction is to motivate participants to be willing to spend the time to use Deduction. One way to manipulate this kind of motivation would be to make the experimental session take a fixed amount of time rather than a fixed number of trials. The instructions to participants could emphasize accuracy and the fact that the participant will gain no advantage from responding quickly.

One of the questions related to motivation is differentiating between participants who cannot use Deduction effectively and those who choose not to. One possible way to increase the use of Deduction would be to inform participants about what response strategy or strategies they will be able to use on a trial. If a participant is explicitly told that they will have the potential to use Deduction, it may encourage them to attempt to use Deduction. In Experiments 1a and 1b, participants were not necessarily able to use Deduction on every trial, which could have led them to focus on easier or faster response strategies. One indication that this method of
motivating participants may not be particularly effective comes from Hardman and Cowan (2016). In that study, all possible response options were displayed on every trial, which means that participants who had even one item in WM were guaranteed to be able to improve their odds of responding correctly by using Deduction. In spite of this, Hardman and Cowan found only moderate use of Deduction although participants were always able to do so. Thus, this approach to motivating participants may be ineffective.

5.5 Response Time Pressure

We found that participants in the reasoning-only task respond more slowly when Deduction is required than when Matching is required. This is true even when only correct responses are considered. This response time difference between Matching and Deduction could allow for the manipulation of the amount of Deduction people are able to use by limiting the amount of time participants are given to respond. This manipulation may selectively hurt Deduction, but not Matching, but even if it only differentially affects relative amounts of Deduction and Matching, it could be useful. This response interval manipulation could be used for the purposes of model validation by varying the amount of time pressure and examining to what extent $\beta$ changes in response.

5.6 Cross-Experiment Parameter Comparisons

To facilitate the comparison of parameter values across experiments, we include the parameter means for Experiments 1a and 1b as well as the feature-matching and change-detection tasks from Hardman and Cowan (2016) in Table 5.1. There are some patterns that stand out.
Accounting for Familiarity with the $\lambda$ parameter appears to reduce $K$ estimates. This makes sense because if a portion of correct responses are accounted for by Familiarity, then those responses cannot also be accounted for by $K$, which in our models is only used for Matching and Deduction. Hardman and Cowan (2016) were unable to account for Familiarity due to limitations in their data set, but there is no reason to think that participants were unable to use Familiarity in that study. As a result, Familiarity that was used by participants but not explicitly modeled may have resulted in inflated $K$ estimates. There are of course also methodological differences between the experiments for which Familiarity was included in the model and the experiments for which it was not included. One example is that Experiment 1a and 1b used partial sample arrays (with some hidden objects) while those of Hardman and Cowan used full sample arrays with no hidden objects. The logical connection between Familiarity and $K$ is, however, quite clear, so we believe that whether or not Familiarity is modeled is likely to have an important impact on estimates of $K$.

The WM capacity parameter, $K$, appears to be higher in change-detection tasks (Experiment 1b and the change-detection task from Hardman & Cowan, 2016) than in feature-matching tasks (Experiment 1a and the feature-matching task from Hardman and Cowan). It is reasonable to interpret this difference as being due to the complexity of the test situation. Change-detection tasks require a decision about a single response option, whereas 2AFC tasks require a choice between two competing response options, which could require both response options to be considered. Although we are attempting to model a number of facets of participant behavior at test, the difference between change-detection and 2AFC tests is one that we choose to neglect for now, although future research may benefit from addressing it.

The Deduction parameter, $\beta$, estimates do not vary in an obvious way with type of response (2AFC vs change-detection) or whether Familiarity was modeled. This stability may be real or it may be an artifact resulting from the wide credible intervals.
Table 5.1: Mean Psychometric Model Parameter Values from Experiments 1a and 1b and Hardman & Cowan (2016)

<table>
<thead>
<tr>
<th>Exp.</th>
<th>$K$</th>
<th>$\beta$</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>1.08</td>
<td>0.50$^a$</td>
<td>0.19$^a$</td>
</tr>
<tr>
<td>1b</td>
<td>1.46</td>
<td>0.49</td>
<td>0.21</td>
</tr>
<tr>
<td>FM$^b$</td>
<td>1.54</td>
<td>0.58</td>
<td>0$^e$</td>
</tr>
<tr>
<td>CD$^c$</td>
<td>2.09</td>
<td>0.51$^d$</td>
<td>0$^e$</td>
</tr>
</tbody>
</table>

$^a$ At array size 3.
$^b$ Feature-matching task from Hardman & Cowan (2016).
$^c$ Change-detection task from Hardman & Cowan (2016).
$^d$ This parameter was called $\alpha$ in Hardman & Cowan (2016).
$^e$ $\lambda$ was not explicitly in the model. If it had been, it would have been set to 0.

of $\beta$. Wide credible intervals for parameters that exist within a bounded range, as $\beta$ does, have the effect of pushing the posterior mean toward the center of the bounded range, which is the value 0.5. As long as $\beta$ has wide credible intervals, it is difficult to obtain $\beta$ values far from 0.5. Note, however, that at array size 1 in Experiment 1a the $\beta$ value was near 1, although array size 1 is an unusual case.

### 5.7 Recall-to-Reject and Familiarity

This research has parallels to research on the recall-to-reject strategy and the use of familiarity in long-term memory tasks. An important strategy involving a kind of deductive reasoning that participants use in long-term memory tasks was identified by Tulving (1983) and is called recall-to-reject (e.g. Gallo, Bell, Beier, & Schacter, 2006) or recollection rejection (e.g. Brainerd, Reyna, Wright, & Mojardin, 2003).

As an example of a paradigm in which recall-to-reject is important, Schmid, Herholz, Brandt, and Buchner (2010) required participants to study words in different semantic categories. In one experiment from that study, a participant might study the word “apple” in the “fruit” category and no other words in the “fruit” category. The participant knows that there was only one studied word in the “fruit” category, so if they are tested with the lure “pear”, they can reject the lure if they recall that they
studied the word "apple" as the exemplar of the fruit category. Thus, recall-to-reject is closely conceptually related to the Deduction process participants might use in our tasks. The only clear difference between recall-to-reject and Deduction is whether the memoranda are stored in long-term memory or WM. This difference, however, may be unrelated to a process that performs the calculations required to reject an invalid lure at test. If the use of Deduction involves the same processes as recall-to-reject, our tasks may provide a WM-based equivalent to long-term memory tasks in which recall-to-reject is used.

In long-term memory research, familiarity typically does not require any information about the context in which an item was presented, only information about the item itself (e.g. Yonelinas, 2001). Our definition of Familiarity differs somewhat from the long-term memory literature definition of familiarity in a few ways. The way in which we have defined Familiarity requires contextual information, in the sense that participants must remember the set of features presented on a single trial. Familiarity does not require the full context, however, which allows the bindings between the colors and orientations of objects to be lost. With respect to knowledge of the context in which items are presented, Familiarity is somewhat different from the long-term memory definition of familiarity. In addition, it is difficult or impossible to use Familiarity if only some of the features are remembered, while using familiarity in long-term memory tasks requires memory of only a single item. Thus, our definition of Familiarity is somewhat different from the definition of familiarity in long-term memory research. Still, our tasks and models in Experiments 1a and 1b begin to bridge the gap between the WM and long-term memory literature on familiarity and recall-to-reject. Future research could examine to what extent recall-to-reject and familiarity correlate with Deduction and Familiarity, which can be measured with our tasks. If such correlations exist, it would suggest commonalities in how participants use strategies to respond in both WM and long-term memory tasks.
5.8 Final Conclusions

We tested three hypotheses about the possible cause of the Deduction Discrepancy. We found evidence against any one of the three hypotheses being a complete explanation for the Deduction Discrepancy. Two of the hypotheses, Hypotheses 1 and 3, provided partial, but decidedly incomplete, explanations for the Deduction Discrepancy. Still, there is some evidence that maintenance of items in WM and additional WM load both impact the use of Deduction to some extent. Our results point toward Deduction being specifically impacted at higher array sizes when WM maintenance of those items is required. We have suggested some hypotheses that could be examined in future research and developed new methods and models that could be used to examine those hypotheses.
Appendix A

Model Details

For the models from both Experiment 1a and 1b, we estimated the parameters using standard Markov chain Monte Carlo techniques. We ran 30,000 iterations and removed 1000 burn-in iterations, leaving 29,000 iterations for use in analysis. The details of model specifications and the priors follow.

A.1 Experiment 1a Model

What follows is a complete model specification for the model used in Experiment 1a. In the following equations, \( C_{ijt} \) is the number of correct responses for the \( i \)th participant, \( j \)th array size and \( t \)th test situation, \( T_{ijt} \) is the number of trials, and \( P_{ijt} \) is the probability of responding correctly. \( A_{ijt} \) is the probability that a participant responds correctly, given that they are paying attention. For convenience, the equations for \( A_{ijt} \) are written in terms of 1 minus the probability of an incorrect response.
\( C_{ijt} \sim \text{Binomial}(T_{ijt}, P_{ijt}) \) \hspace{1cm} (A.1)

\( P_{ijt} = aA_{ijt} + (1 - a)(1/2) \) \hspace{1cm} (A.2)

\( d_1 = \min(K_i/N_j, 1) \) \hspace{1cm} (A.3)

\( d_2 = \min(K_i/(N_j - 1), 1) \) \hspace{1cm} (A.4)

\( A_{ij1} = 1 - (1 - d_1)[\beta_{ij} * (1 - d_2) + (1 - \beta_{ij})](1/2) \) \hspace{1cm} (A.5)

\( A_{ij2} = 1 - (1 - d_1)(1 - \lambda_{ij})(1/2) \) \hspace{1cm} (A.6)

\( A_{ij3} = 1 - [\beta_{ij} * (1 - d_1) + (1 - \beta_{ij})](1 - \lambda_{ij})(1/2) \) \hspace{1cm} (A.7)

Note that the attention parameter, \( a \), was set to the constant value 0.95. The \( \beta \) and \( \lambda \) parameters were allowed to vary with array size by adding an array size main effect (e.g. \( \beta'_j \)) to participant parameters (e.g. \( \beta'_i \)), as is shown in the following equations. To constrain parameter estimation, we constrained \( \beta'_j \) and \( \lambda'_j \) at array size 4 to be 0.

\( K_i = \max(K'_i, 0) \) \hspace{1cm} (A.8)

\( K'_i \sim \text{Normal}(\mu_K, \sigma^2_K) \) \hspace{1cm} (A.9)

\( \beta_{ij} = LT(\beta'_i + \beta'_j) \) \hspace{1cm} (A.10)

\( \beta'_i \sim \text{Normal}(\mu_\beta, \sigma^2_\beta) \) \hspace{1cm} (A.11)

\( \beta'_j \sim \text{Cauchy}(0, 1) \) \hspace{1cm} (A.12)

\( \lambda_{ij} = LT(\lambda'_i + \lambda'_j) \) \hspace{1cm} (A.13)

\( \lambda'_i \sim \text{Normal}(\mu_\lambda, \sigma^2_\lambda) \) \hspace{1cm} (A.14)

\( \lambda'_j \sim \text{Cauchy}(0, 1) \) \hspace{1cm} (A.15)
\( LT \) is the logit transformation, the CDF of the standard logistic distribution which transforms from the unbounded space to the interval \([0, 1]\). The hyperparameters had the following fixed priors.

\[
\begin{align*}
\mu_{K'} &\sim \text{Normal}(1.5, 1) \quad (A.16) \\
\sigma^2_{K'} &\sim \text{Inverse Gamma}(0.1, 0.1) \quad (A.17) \\
\mu_{\beta'} &\sim \text{Normal}(0, 2) \quad (A.18) \\
\sigma^2_{\beta'} &\sim \text{Inverse Gamma}(0.1, 0.1) \quad (A.19) \\
\mu_{\lambda'} &\sim \text{Normal}(0, 2) \quad (A.20) \\
\sigma^2_{\lambda'} &\sim \text{Inverse Gamma}(0.1, 0.1) \quad (A.21)
\end{align*}
\]

Note that the prior means on \( \beta'_i \) and \( \lambda'_i \), \( \mu_{\beta} \) and \( \mu_{\lambda} \), have priors centered on 0. However, once transformed with \( LT \), 0 becomes 0.5. Thus, the priors on \( \beta_i \) and \( \lambda_i \) are centered on moderate probabilities.

### A.2 Experiment 1b Model

In the model for Experiment 1b, \( S_{it} \) is the number of Same responses for the \( i \)th participant and \( t \)th test situation, \( T_{it} \) is the number of trials, and \( P_{it} \) is the probability of a Same response (i.e. the feature response). \( D_{it} \) is the probability that participant \( i \) gives a Different response (i.e. the rejection response) given that the participant is paying attention.
\begin{align*}
S_{it} & \sim \text{Binomial}(T_{it}, P_{it}) \quad (A.22) \\
P_{it} & = a(1 - D_{iit}) + (1 - a)(1/2) \quad (A.23) \\
d_1 & = \min(K_i/N, 1) \quad (A.24) \\
d_2 & = \min(K_i/(N - 1), 1) \quad (A.25) \\
D_{i1} & = (1 - d_1)(1 - \lambda_i)(1 - g_i) \quad (A.26) \\
D_{i2} & = 1 - [\beta_i(1 - d_1)(1 - d_2) + (1 - \beta_i)][\lambda + (1 - \lambda_i)g_i] \quad (A.27) \\
D_{i3} & = 1 - [\beta_i(1 - d_1) + (1 - \beta_i)](1 - \lambda_i)g_i \quad (A.28) \\
D_{i4} & = D_{i3} \quad (A.29) \\
D_{i5} & = 1 - g_i \quad (A.30)
\end{align*}

For this experiment, \( a = 0.95 \) and \( N = 3 \). The priors and transformations were all the same as for Experiment 1a with the following exceptions: (a) There were no effects of array size as only one array size was used and (b) the additional \( g_i \) parameters have the given prior.
\[ K_i = \max(K_i', 0) \quad (A.31) \]

\[ K_i' \sim \text{Normal} (\mu_K, \sigma_K^2) \quad (A.32) \]

\[ \beta_i = LT(\beta_i') \quad (A.33) \]

\[ \beta_i' \sim \text{Normal} (\mu_\beta, \sigma_\beta^2) \quad (A.34) \]

\[ \lambda_i = LT(\lambda_i') \quad (A.35) \]

\[ \lambda_i' \sim \text{Normal} (\mu_\lambda, \sigma_\lambda^2) \quad (A.36) \]

\[ g_i = LT(g_i') \quad (A.37) \]

\[ g_i' \sim \text{Normal} (\mu_g, \sigma_g^2) \quad (A.38) \]

The hyperparameters had the following fixed priors. The prior on \( \mu_K \) was Normal(1.5, 1). The priors on \( \mu_\beta \), \( \mu_\lambda \), and \( \mu_g \) were Normal(0, 2). The priors on \( \sigma_K^2 \), \( \sigma_\beta^2 \), \( \sigma_\lambda^2 \), and \( \sigma_g^2 \) were Inverse Gamma(0.1, 0.1).
Appendix B

Shrinking $\beta$ Credible Intervals

We have consistently found wide credible intervals for the $\beta$ parameter, which means that we do not have clear information about the value of the parameter. This is problematic for reasons discussed in the main text. To reduce the width of the credible interval of $\beta$, we used some variants of the model for the Experiment 1a task.

In the first variant, we constrained which array sizes were allowed to differ from one another. In the full model (the best model selected in the main text), we found that $\beta$ did not differ between array sizes 3 and 4 and that $\lambda$ did not differ between array sizes 1 and 3. Thus, it may be possible to reduce the width of the credible intervals of $\beta$ by forcing $\beta$ to have to same value at array sizes 3 and 4 and forcing $\lambda$ to have the same value at array sizes 1 and 3.

In the second variant, we used the constraints of the first variant and also constrained $\beta$ to have the same value for all participants. This is a strong constraint with the unfortunate side effect of losing any information about individual differences between participants.

For the full model that was used in the main text, the credible interval for the mean $\beta$ at array size 3 was (0.16, 0.89), which is 0.73 wide. For the first variant we fit here, the credible interval was (0.29, 0.86), which is 0.57 wide. For the second
variant, the credible interval was (0.44, 0.65), which is 0.21 wide. Thus, constraining
the $\beta$ and $\lambda$ parameters produce narrower credible intervals. But what is the cost in
terms of model fit?

The WAIC values were 1233 for the full model, 1228 for the first variant, and
1260 for the second variant. The first variant actually fits better than the full model,
while the second variant fits somewhat worse. Although the fit of the second variant
is worse, it does allow for smaller credible intervals than either the full model or the
first variant. Thus, greater specificity about the mean $\beta$ value can be obtained by
sacrificing some model fit and information about individual participant $\beta$ parameters.
Note, however, that individual $\beta$ parameters have very wide credible intervals, so very
little information about those parameters is available even when they are included in
the model.
Appendix C

Simplified Experiment 1a Model

The data from Experiment 1a can be analyzed with a model that is much more simple than the model we used in the main text. The more simple model has the advantage that its parameters can be calculated with closed form equations. As we will show, the parameters from the simple model correlate well with parameters from the model from the main text. The main downside of the simple model is the lack of hierarchical constraint afforded by the Bayesian primary model. In addition, the simple model cannot use data from array size 1 because it requires that all 3 test situations to be present at each set size.

The model is presented in Figure C.1. Beginning from the start node S, the participant successfully uses Matching, Deduction, or Familiarity, the probability of each being given by the parameters $M$, $D$, and $F$, respectively. Thus, while $\beta$ represents an attempt to use Deduction, $D$ represents the success of Deduction. Note that, depending on which test situation a participant is in, they are unable to use one of the three memory-based response strategies. Thus, for any given test situation, the general model tree is simplified by one of $M$, $D$, or $F$ being 0, which effectively removes that branch from the model.

A guess occurs whenever all three memory-based strategies fail. $M$, $D$, and $F$ are
Figure C.1: Binomial process trees for the Experiment 1a simplified model. Beginning from the start node, marked S, the participant can use Matching, Deduction, or Familiarity to respond correctly. Note that each test situation prevents the use of one response strategy, which means that $F$, $D$, and $M$ are 0 in test situations 1, 2, and 3, respectively. If all memory-based response strategies fail, which happens with probability $G$, the participant enters a guessing state and responds correctly half of the time.
the rates at which each memory-based strategy succeeds. Thus, the rate at which
guessing occurs, $G$, is given by

$$G = 1 - (M + D + F).$$

The rate at which a guess does not occur, $1 - G$, is given by

$$1 - G = M + D + F.$$

An incorrect response only happens when all three memory-based strategies fail and
the guess is incorrect. This happens with probability $G \times 1/2$. A correct response
happens whenever the response is not incorrect. Thus, the correct response rate, $PC$, is
given by

$$PC = 1 - G \times 1/2.$$ 

We can solve for a general form of $PC$ in terms of the parameters by substituting in
$1 - (M + D + F)$ for $G$.

$$PC = 1 - 1/2 \times (1 - (M + D + F)) \quad (C.1)$$
$$= 1 - 1/2 + 1/2 \times (M + D + F) \quad (C.2)$$
$$= 1/2 + 1/2 \times (M + D + F) \quad (C.3)$$
$$= 1/2 \times (1 + M + D + F) \quad (C.4)$$

For a given test situation, one of the response strategies is impossible, so the $G$
for each test situation are simplified by one of the parameters being 0.

\[ G_1 = 1 - M - D \]  \hspace{1cm} (C.5)
\[ G_2 = 1 - M - F \]  \hspace{1cm} (C.6)
\[ G_3 = 1 - D - F \]  \hspace{1cm} (C.7)

As a result, the \( PC \) for each test situation are simplified as follows,

\[ PC_1 = 1/2 \times (1 + M + D) \]  \hspace{1cm} (C.8)
\[ PC_2 = 1/2 \times (1 + M + F) \]  \hspace{1cm} (C.9)
\[ PC_3 = 1/2 \times (1 + D + F) \]  \hspace{1cm} (C.10)

We have the three equations for \( PC \) for each test situation and these equations contain three unknowns. Thus, we can solve for the three parameters. We can solve for \( M \) by adding equations containing \( M \) and subtracting equations not containing \( M \). Then both sides get multiplied by 2. Like terms are collected, which cancels terms containing the parameters other than \( M \). Finally, the equation is rearranged to solve for \( M \).

\[ PC_1 + PC_2 - PC_3 = 1/2(1 + M + D) + 1/2(1 + M + F) - 1/2(1 + D + F) \]  \hspace{1cm} (C.11)
\[ 2PC_1 + 2PC_2 - 2PC_3 = 1 + M + D + 1 + M + F - 1 - D - F \]  \hspace{1cm} (C.12)
\[ 2PC_1 + 2PC_2 - 2PC_3 = 1 + 2M \]  \hspace{1cm} (C.13)
\[ M = PC_1 + PC_2 - PC_3 - 1/2 \]  \hspace{1cm} (C.14)
Table C.1: Primary Model and Simple Model Parameter Correlations

<table>
<thead>
<tr>
<th></th>
<th>$K$</th>
<th>$\beta$</th>
<th>$\lambda$</th>
<th>$M$</th>
<th>$D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td></td>
<td>0.57*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.46*</td>
<td>0.53*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M$</td>
<td>0.74*</td>
<td>0.33</td>
<td>0.16†</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D$</td>
<td>0.70*</td>
<td>0.89*</td>
<td>0.55*</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>$F$</td>
<td>0.33</td>
<td>0.32</td>
<td>0.92*</td>
<td>0.00†</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Correlation coefficients that are reliably nonzero are marked with * while those that are reliably zero are marked with †.

This result has a nice pattern, in that $PC_1$ and $PC_2$, which both depend on $M$, are added, while $PC_3$, which does not depend on $M$, is subtracted. This pattern holds for the other parameters of the model and the full set of solutions is

\[
M = +PC_1 + PC_2 - PC_3 - 1/2 \quad \text{(C.15)}
\]

\[
D = +PC_1 - PC_2 + PC_3 - 1/2 \quad \text{(C.16)}
\]

\[
F = -PC_1 + PC_2 + PC_3 - 1/2. \quad \text{(C.17)}
\]

Note that none of the parameters can be less than 0. In addition, the sum of any pair of parameters cannot be greater than 1. These two constraints are not imposed by the equations, so they should be tested for and corrected after calculating the parameters.

We calculated the simple model parameters at array sizes 3 and 4, then averaged those parameter values. We then correlated the primary model and simple model parameters. These correlations are presented in Table C.1.

One result that stands out is that $K$ is correlated with both $M$ and $D$. This makes sense because both the Matching and Deduction strategies use WM capacity. It highlights, however, a limitation of the simple model, which is that it does not cleanly separate capacity and Deduction, which are estimated with $K$ and $\beta$, respectively, in the primary model. In spite of this, $\beta$ and $D$ are correlated 0.87 and $\lambda$ and $F$...
are correlated 0.92, which indicates that those pairs of parameters are capturing very similar information. In addition, \( K \) and \( M \) are correlated 0.73.

In general, the parameters from the simple model are highly correlated with parameters from the primary model used in the main text. For many purposes, it may be sufficient to use the simple model, which is much easier to use than the primary model.
References


Cowan, N., Blume, C., & Saults, S. (2013). Attention to attributes and objects in working memory. Journal of Experimental Psychology: Learning, Memory, and


Hardman, K. O. (2016). CX (version 0.2.2) [Computer software]. Retrieved from https://github.com/hardmanko/ofxCX/releases/tag/v0.2.2

Hardman, K. O. (2017). CMBBHT: Cell means based Bayesian hypothesis tests (version 0.1.2) [Computer software]. Retrieved from https://github.com/hardmanko/CMBBHT/releases/tag/v0.1.2


Ricker, T. J., & Cowan, N. (2010). Loss of visual working memory within seconds:


Kyle Hardman was born in Ruch, Oregon. He graduated in 2010 with a bachelor of sciences in psychology from Southern Oregon University. In 2011, he enrolled in the University of Missouri doctoral program in the Department of Psychological Sciences with Nelson Cowan as his advisor. In 2013, Kyle received a master of arts in Psychology from the University of Missouri for a thesis titled *Remembering Complex Objects in Visual Working Memory*. Kyle continued working with Nelson Cowan for the remainder of his doctoral program, completing the degree in 2017 with a dissertation titled *Reasoning and Memory: Multiple Simple Response Strategies are used in Visual Working Memory for Color-Orientation Binding*. 