1 Experiment 1

The goals for this experiment were twofold. First, I hoped to test order effects in a paradigm built for a dual-systems model of categorization. Despite the fact that many approaches posit such a model, only one approach (COVIS) has explicitly tested how a given individual switches between the two systems for categorization. Most other studies that compare different types of categorization do so in a between-subjects manner (e.g., **??**, described above). Thus, this study is one of the first to look at the interaction between two categorization systems within-subjects.

This experiment also has a more practical purpose. One of the overarching goals of this dissertation is to compare different paradigms of category learning. However, all studies using a statistical density approach have been done between subjects, so we do not yet know if there are any transfer effects for this type of task. This experiment carefully conducts three order analyses to fully understand the Sloutsky-style category learning task. This will benefit both practical understanding of this task as well as its underlying theoretical framework.

An interesting feature to the task used here and in **?** is the two manipulations, each of which engages or requires one of the two category learning systems. First, we have the instruction type. Differing task demands engage either the associative or the hypothesis-testing system. The second manipulation is stimulus type. Dense and sparse stimuli are best learned by different systems. Thus, each block can either be a match (where the instruction type engages the ideal system for the stimulus type) or a mismatch (where the instruction type engages the wrong system for the stimulus type).

To fully understand this task, I test for order effects in both matching and mismatching conditions. First, I tested order for the matching conditions. This analysis is the most simple and straightforward test of order; can participants switch between systems when both the stimulus type and the learning type cue a certain system? The second order analysis tested dense stimuli, looking to see whether participants were able to engage the associative system even when the learning instructions sometimes cued the hypothesis-testing system. Further, if participants were not able to overcome supervised learning instructions and ended up using the hypothesis-testing system in their first block, this analysis investigated whether they could switch to the associative system in subsequent blocks. Finally, the third order analysis tested sparse stimuli. Similar to the prior order analysis, here I tested whether participants could engage the hypothesis-testing system to learn sparse stimuli even when task demands cued the associative system. This order analysis also tested the participants' ability to switch away from the associative system.

1.1 Method

1.1.1 Participants

Data was collected from 236 undergraduate psychology students at the University of Connecticut (161 Female, 67 Male, mean age = 18.94). Data for the category learning task was lost for 7 subjects due to technical errors. Thus, the final sample size was 229. Each subject was placed into one of six groups. Each group completed two blocks of the category learning task in a specific order. For more details, see Table 3. Unequal group sizes result from lost data due to technical errors.

Table 1

Group sizes for each order.

Analysis	Group	*N*
1	1	40
	2	38
2	3	39
	4	39
3	5	36
	6	37

1.1.2 Category Learning Task

This task measures learning of dense and sparse categories and is based off of a paradigm from previous research (**?**). Participants learn novel categories of items in four possible conditions in a 2 x 2 design. The first manipulation is learning type (supervised vs. unsupervised). In *supervised* learning, participants learn the categories by being instructed on the relevant features (e.g., "All friendly aliens have big noses."). Images of the relevant features are provided along with the descriptions. In *unsupervised* learning, participants learn the categories by viewing sixteen instances of the category.

Table 2

Relationship between learning systems and experimental manipulations.

Experimental feature	Hypothesis-testing	Associative
Learning type	Supervised	Unsupervised
Stimulus type	Sparse	Dense

The second manipulation is category type (sparse vs. dense). Category type is measured by statistical density, which ranges from zero (where all features vary freely) to one (where all features co-occur perfectly). It is based on a comparison between within- and between-category entropy (**?**). All categories in this experiment have seven dimensions. The *sparse* categories cohere on a single dimension, while the other dimensions vary freely (density = .25). In contrast, the *dense* categories cohere on six of the seven dimensions (density = .75). The seventh dimension is allowed to vary freely. For more details on how density was calculated, see Appendix A. Stimuli for each of the four blocks are different. See Fig. 1 for examples of the experimental manipulations. Table 2 summarizes how each experimental manipulation corresponds to the theorized category learning systems.

This task is within-subjects. Based on the group they were placed into, participants completed two of the four possible learning-category type combinations. In this experiment, I conducted three different order analyses. This design led to six possible order groups that each participant could be placed into. See Table 3 for a summary.

In each block, participants were introduced to the task through a short cover story. They were told to learn which items go with a certain property (e.g., which aliens are friendly). Crucially, no labels were attached to the categories (e.g., some aliens are Ziblets). Then, participants completed a training block (either supervised or unsupervised). During training, only members of the target category or its features were shown. After training, participants completed 40 test trials (16 target, 16 distractor, 8 catch), following the design of **?**. In each trial, participants saw a single item and used the keyboard to indicate whether the item matched the category they had just learned (e.g., if the alien is friendly). Catch items looked significantly different than both the target and competing categories, so participants should have always rejected them as members of the learned category. This experiment was presented using PsychoPy v.1.84.2 (**?**).

Table 3

Block orders for statistical density task.

Analysis	Group	Block 1	Block 2
1	1	Unsupervised-dense	Supervised-sparse
	2	Supervised-sparse	Unsupervised-dense
2	3	Unsupervised-dense	Supervised-dense
	4	Supervised-dense	Unsupervised-dense
3	5	Unsupervised-sparse	Supervised-sparse
	6	Supervised-sparse	Unsupervised-sparse

1.1.3 Behavioral Measures

I used multiple assessments to test participants' language ability. The choice of assessments was based on the epiSLI criteria for language impairment (**?**), which includes comprehension, expression, vocabulary, grammar, and narrative. I adapted these requirements from a kindergarten population to a college-aged population. The epiSLI criteria have been shown to be robust for diagnosis of specific language impairment (SLI). In addition, other studies of language impairment more broadly have adapted a similar multidimensional approach to measuring language ability, sometimes including measures of phonological skills (**?**). Thus, using assessments that cover the many domains of language outlined in epiSLI criteria will allow me to get a fuller picture of individual differences in language ability. See Table 4 for a summary of the assessments and which domains of the epiSLI criteria they cover. The specific tests used in this experiment are detailed below.

Test of word reading efficiency (TOWRE) phonemic decoding subtest. TOWRE is a test of nonword fluency (**?**). This test is a part of the comprehension aspect of epiSLI, since the comprehension measure is reading-based. In the TOWRE, individuals have 45 seconds to read as many nonwords as possible. The nonwords become longer and more difficult as the list goes on. The raw score from the TOWRE is calculated by counting the number of words correctly pronounced before the time limit. These raw scores are then converted to standard scores using age-based norms. The standard scores are based on a distribution with a mean of 100 and a standard deviation of 15. In the current age range, a perfect raw score (63) on the TOWRE returns a standard score of ">120." For the purposes of this study, scores of ">120" will be trimmed to 120.

Woodcock Johnson-III word attack (WA) subtest. This task measures nonword decoding ability (**?**). Like the TOWRE, it is helpful for measuring the comprehension aspect of epiSLI. However, while the TOWRE measures word fluency, this task measures decoding accuracy. Participants read a list of nonwords out loud at their own pace. Raw scores are calculated by counting the number of words the participant said correctly. Raw scores are converted to standard scores using age-based norms. The standard score distribution has a mean of 100 and a standard deviation of 15.

Computerized reading comprehension. This test covers the comprehension and narrative aspects of epiSLI. This computerized reading comprehension (CRC) test is based on the Kaufman Test of Educational Achievement (KTEA) reading comprehension subtest (**?**). To create this test, I copied the passages and questions contained in the KTEA reading comprehension subtest into E-Prime (**?**) for presentation on a computer. Then, I created multiple choice answers for the KTEA questions that did not already have them. In this task, participants read short expository and narrative texts and answer multiple-choice comprehension questions about them. Some questions are literal, while others require participants to make an inference. Participants completed as many questions as they could in 10 minutes. Once 10 minutes had elapsed, the participant was allowed to answer the question currently on the screen and then the assessment closed. Because this task is a modified version of the KTEA, I use raw scores in analysis rather than standardized scores based on the KTEA norms. Raw scores are calculated by counting the number of correctly answered questions for each participant.

Nelson-Denny vocabulary subtest. The Nelson-Denny vocabulary sub-test is a written assessment of vocabulary (**?**). This test covers the vocabulary aspect of epiSLI. This test has been used in multiple studies of college-aged adults and provides sufficient variability for individual difference investigations in this population (e.g., **?**; **?**). In this test, participants are asked to choose the word closest to a target vocabulary word. The test has a total of 80 items. Participants were allowed unlimited time to complete all items. Raw scores were generated by counting the total number of correctly answered items. The raw scores were then converted to standard scores based upon a norming sample including students in 10th, 11th, and 12th grade as well as two- and four-year college students. The standard scores for this assessment have a mean of 200 and a standard deviation of 25.

Clinical Evaluation of Language Fundamentals recalling sentences subtest. I used the Recalling Sentences subtest from the Clinical Evaluation of Language Fundamentals - Fourth Edition (CELF; **?**) to cover the grammar and expression aspects of epiSLI. In this subtest, participants hear sentences and are asked to repeat them. Scoring is based on how many errors the participant makes in their repetition. Raw scores are calculated by adding up the number of points achieved for each item. These are then converted to standard scores using age-based norms. The standard scores are based on a distribution with a mean of 10 and a standard deviation of 3.

Table 4

Assessments of language and their corresponding epiSLI domains.

Test	epiSLI Criteria
TOWRE WA	Comprehension (decoding aspect)
CRC	Comprehension, narrative
ND Vocab	Vocabulary
CELF RS	Grammar, expression

Raven's Advanced Matrices. Finally, I used Set II of Raven's Advanced Matrices (RAM) to measure nonverbal IQ (**?**). In this task, participants see a grid containing eight images and an empty space. The images are arranged in the grid according to some rule or rules. Participants must choose one of eight additional images that fits in the empty space. Due to time constraints, I restricted participants to 10 minutes in this task. Since this administration is different than the standard administration, I do not use standard scores. Raw scores are calculated by counting the number of correct answers given within 10 minutes.

1.2 Procedure

Each participant completed the category learning task as well as all of the behavioral measures. TOWRE, WA, and CELF were audio-recorded to allow for offline scoring. To allow multiple subjects to be run in a single timeslot, some participants received tasks they could complete on their own (category learning, ND, computerized reading comprehension, Raven's) first while others completed tasks with the experimenter first (WA, CELF, TOWRE). All together, the seven tasks took approximately one hour.

1.3 Results

For all analyses shown below, accuracy was converted to *d'* values (**?**) using the R package **neuropsychology** (**?**). Correction for extreme values was done following (**?**). Following prior research, all blocks where 5 or fewer catch items were correctly rejected were dropped from analysis. This resulted in 22 total missing blocks (out of 458 total), including both blocks from a single subject in group 5. For basic descriptive statistics on the category learning task, see Table 5. For reaction time, all incorrect trials were dropped, as well as any trials with a response faster than 250ms.

Table 5

Descriptive statistics for the category learning task.

Analysis	Group	Block	Mean (SD) Accuracy	Mean (SD) RT (ms)
1	1	Unsupervised-dense	0.91 (0.19)	1074 (697)
		Supervised-sparse	0.93 (0.14)	809 (549)
	2	Supervised-sparse	0.72 (0.34)	864 (608)
		Unsupervised-dense	0.90 (0.18)	834 (563)
2	3	Unsupervised-dense	0.91 (0.18)	1063 (674)
		Supervised-dense	0.91 (0.23)	946 (636)
	4	Supervised-dense	0.90 (0.21)	960 (679)
		Unsupervised-dense	0.92 (0.18)	861 (561)
3	5	Unsupervised-sparse	0.57 (0.35)	1275 (761)
		Supervised-sparse	0.93 (0.14)	812 (543)
	6	Supervised-sparse	0.93 (0.12)	866 (592)
		Unsupervised-sparse	0.53 (0.38)	1003 (633)

1.3.1 Behavioral Assessments

For basic descriptive statistics on the behavioral measures, see Table 6. Before performing any statistical analyses using the individual difference measures, I checked their normality using the D'Agostino normality test from the R package **fBasics** (**?**). Four measures (CRC, ND Vocab, CELF RS, RAM) were significantly skewed. These measures were centered, scaled, and transformed using Yeo-Johnson transformations from R package **caret** (**?**). The remaining measures (TOWRE, WA) were not skewed and thus were simply scaled and centered.

Table 6

Descriptive Statistics for Behavioral Measures

Assessment	Mean	SD	Range
CELF Recalling Sentences SS	10.7	1.86	3-14
Computerized Reading Comprehension	21.7	5.12	7-48
Nelson-Denny Vocabulary SS	229	14.0	175-255
TOWRE SS	96.2	9.86	59-120
Word Attack SS	99.7	9.04	75-120
Raven's Advanced Matrices	15.1	4.58	0-26

Since my goal was to create a composite measure of language ability, I investigated the relationship between the behavioral measures. First, I constructed a correlation matrix between all of the behavioral measures (see Table 7). All pairs of measures had a significant positve correlation with the exception of CELF RS and RAM. To further test whether the behavioral measures could be combined into a single composite, I ran a principal components analysis (PCA) on the 5 assessments related to epiSLI (i.e., all assessments except RAM). The Kaiser-Meyer-Olkin overall measure of sampling adequacy was 0.69, above the commonly accepted threshold of 0.6. Bartlett's test of sphericity was also significant $\chi^2(10)$ = 236.16, *p* $<$0.001. These suggest that the 5 behavioral assessments were suitable for a PCA.

The first component in the PCA accounted for 47.74% of the variance and had an eigenvalue of 2.38. All of the factor loadings for this component were quite similar, ranging from -0.41 to -0.51. The second factor accounted for an additional 20.5% of the variance and had an eigenvalue of 1.02. This factor separated the two measures involved in decoding (TOWRE and WA) from the other measures (CRC, ND Vocab, and CELF RS). The remaining components had eigenvalues below 1. Thus, of the two significant components, the first component explained almost half of the variance and had an eigenvalue more than double the second component, which largely represented decoding ability. Since the first component indicated that most of the measures loaded similarly, I decided to take a simple means approach to creating a language composite measure.

The language composite measure was created by averaging the 5 scaled, centered, and/or transformed measures. For participants with missing behavioral measures, the composite was created by averaging the remaining available measures. No subject was missing more than 1 measure. This composite measure was then scaled but not centered. This language composite measure and the centered, scaled, and transformed RAM measure are used in the analyses investigating order effects reported below.

Table 7

Correlations between behavioral measures.

	1	2	3	4	5	6
1. Computerized Reading Comprehension	-					
2. Nelson-Denny Vocabulary	0.57***	-				
3. CELF Recalling Sentences	0.31***	0.40***	-			
4. Raven's Advanced Matrices	0.31***	0.34***	0.09	-		
5. TOWRE	0.22**	0.28***	0.26***	0.16***	-	
6. Word Attack	0.22**	0.38***	0.29***	0.22***	0.53***	-

**p* $<$0.05, ***p* $<$0.001, ****p* $<$0.0001

1.3.2 Order Analysis 1: Matching Conditions

The first analysis investigated order effects for blocks in which the learning type (supervised vs. unsupervised) and category type (sparse vs. dense) both engaged the same category learning system (hypothesis testing vs. associative). Participants completed supervised-sparse and unsupervised-dense blocks.

Accuracy. I used linear mixed-effects models to examine the effects of block and order on accuracy at test. Accuracy in these models was measured by *d'* values for each subject by block. The base model included random intercepts for subject. Adding block and order as fixed effects significantly increased model fit, $\chi^2(2)$ = 13.21, p = 0.001. Adding the interaction between block and order further improved model fit, $\chi^2(1)$ = 6.03, p = 0.014. Thus, the final model including only experimental conditions had fixed effects of block, order, and the interaction between block and order as well as random intercepts for subject.

This model revealed two significant effects. First, there was a significant main effect of order, $F(1,75)$ = 8.60, p = 0.004. There was also a significant interaction between block and order, $F(1,74)$ = 6.10, p = 0.02. There was not a significant main effect of block, $F(1,76)$ = 2.61, p = 0.11. The interaction was broken down by conducting two separate models for each of the orders (unsupervised-dense first and supervised-sparse first). These analyses showed that when the associative system was engaged first (unsupervised-dense first), there was no significant main effect of block, $F(1,36)$ = 0.014, p = 0.91. When the hypothesis testing system was used first (supervised-sparse first), there was a significant effect of block, $F(1,37)$ = 7.52, p = 0.009. This shows that when participants complete engage the hypothesis-testing system first, performance on the supervised-sparse (hypothesis-testing) block is lower than in the unsupervised-dense (associative) block (see Table 5 Figure 2).

To investigate the effect of individual differences in language ability on the order effect, I used the final model above which included main effects for block and order as well as their interaction. I then added the language composite measure as a fixed effect. I also added RAM to control for nonverbal IQ. This model revealed no significant effects for RAM or the language composite; there remained a significant interaction between block and order.

Reaction time. Again, I used linear-mixed effects models to look at the effects of block and order on reaction time at test. While the accuracy measure was at the block level, reaction time here is modeled at the item level. The base model included random intercepts for subject and for block nested within subject. Adding the fixed effects of block and order increased the model fit, $\chi^2(1)$ = 21,02, $p <$0.001. Further, adding the interaction between block and order improved model fit, $\chi^2(1)$ = 29.6, $p <$0.001.

This model showed three significant effects. There was a significant main effect of block, $F(1,72)$ = 52.42, $p <$0.001. There was also a significant main effect of order, $F(1,77)$ = 4.67, p = 0.03. Finally, there was a significant interaction between block and order, $F(1,72)$ = 35.0, $p <$0.001. To break down this interaction, I ran follow-up models for each of the two orders. This showed that when the associative system was engaged first (unsupervised-dense first), there was a significant main effect of block, $F(1,37)$ = 53.6, $p <$0.001. When the hypothesis testing system was used first (supervised-sparse first), there was no significant effect of block, $F(1,35)$ = 0.30, p = 0.59. This result is the opposite of what was found in accuracy. When the associative system is engaged first, we see a difference in RT between blocks, but when the hypothesis-testing system in engaged first, there is no difference in RT.

Similar to the accuracy analysis, I added RAM and language ability as fixed effects to the final model from above. Neither one had any effect on RT. The main effects and interactions from above stayed significant.

Summary. While the findings from accuracy and reaction time seem to be opposing, they may in fact tell the same story. Group 1 engaged the associative system first. This group showed similar accuracy for both blocks but slower reaction time in their first block (unsupervised-dense/associative). Group 2 engaged the hypothesis-testing system first. They showed similar reaction times for both blocks, but lower accuracy in their first block (supervised-sparse/hypothesis-testing). Thus, both groups showed reduced performance (reflected in either reaction time or accuracy) on their first block, regardless of which system it engaged, perhaps reflecting a general learning effect across the task as a whole. Importantly, this learning effect is not modulated by language ability.

1.3.3 Order Analysis 2: Dense Stimuli

The second order analysis compared groups 3 and 4. All participants learned only dense categories, with the order of learning types differing between groups.

Accuracy. Again, I used linear-mixed effects models to investigate the effects of block and order on accuracy at test. The base model included random intercepts for subject. Adding the fixed effects to the model did not significantly improve fit $\chi^2(2)$ = 0.07, p = 0.97. Indeed, neither block, F(1,145) = 0.053, p = 0.82, nor order, F(1,145) = 0.016, p = 0.90, were significant predictors of accuracy. Thus, accuracy at test on dense categories was similar regardless of training type or block order. Next, I conducted the individual differences analysis. Since the goal of this investigation was to see whether the relationship between order and accuracy in each block changed as a function of language ability, I created a model with fixed effects for block, order, and language ability as well as RAM. The model showed no significant effects of any of the predictors.

Reaction time. I used the same linear mixed-effects model as above, with random intercepts for subject and for block nested within subject in the base model. Adding fixed effects of order and block did not significantly improve model fit, $\chi^2(2)$ = 2.54, p = 0.28. Block, F(1,72) = 0.05, p = 0.83, and order, F(1,77) = 2.49, p = 0.12, did not have any effect on reaction time. Adding language ability and RAM to the model also did not improve fit, $\chi^2(2)$ = 2.46, p = 0.12. These measures were not significant predictors of reaction time for dense stimuli.

Summary. There were no significant effects of block, order, or language ability found for dense stimuli. This may suggest that learning dense stimuli engages a single system regardless of the instructions. Alternatively, it may be that learning dense stimuli is overall an easy task, evidenced by the high accuracy values seen in these blocks.

1.3.4 Order Analysis 3: Sparse Stimuli

The third order analysis investigated differences in learning sparse categories based on learning type order, using data from groups 5 and 6.

Accuracy. I used the same type of linear mixed-effect models as the prior two order effects, with random intercepts for subject. Adding block and order significantly increased model fit, $\chi^2(2)$ = 57.5, $p < 0.001$. However, adding the interaction between block and order did not increase model fit, $\chi^2(2)$ = 0.33, p = 0.56. Thus, the final model included fixed effects for order and block but not their interaction. This model revealed a significant main effect of block, $F(1,67)$ = 75.69, $p < 0.0001$, but no significant main effect of order, $F(1,67)$ = 0.0008, p = 0.98. Participants showed significantly higher accuracy in supervised-sparse blocks than in unsupervised-sparse blocks (see Table 5).

As in the two previous analyses, I added RAM and language ability to the final model above. Adding the language composite improved model fit even after adding RAM, $\chi^2(2)$ = 5.34, p = 0.02. However, adding the interactions between block and language and order and language did not improve model fit, $\chi^2(2)$ = 1.94, p = 0.38. The final model, which included no interactions, showed the same main effect of block seen above as well as a significant main effect of language ability, $F(1,63)$ = 5.21, p = 0.03. The effect of language ability was associated with a positive coefficient, b = 0.19, suggesting that accuracy and language ability were positively related. There was no main effect of RAM.

Reaction time. As above, I used a linear mixed-effect model with random intercepts for subject and block nested within subject as the base model. Adding the fixed effects of order and block significantly improved fit, $\chi^2(2)$ = 50.0, $p < 0.0001$. In addition, adding the interaction between block and order improved fit, $\chi^2(2)$ = 25.4, $p < 0.0001$. The final model showed significant main effects for block, $F(1,67)$ = 39.46, $p < 0.0001$, order, $F(1,70)$ = 4.56, p = 0.04, as well as a significant interaction between block and order, $F(1,67)$ = 29.57, $p < 0.0001$. Follow-up models showed that for each order, there was a significant difference in reaction time by block. However, the difference between mean reaction time of the two blocks for group 5 (unsupervised-sparse first) was 463 ms, while the difference for group 6 (supervised-sparse first) was 137 ms. This suggests that the interaction represents a greater difference between blocks for participants who received the unsupervised-sparse block first.

For the individual differences analysis, I added RAM and the language composite to the final model from above. Adding the language composite did not improve the model fit. There was no effect of language ability on reaction time for sparse stimuli.

Summary. In terms of accuracy, participants showed higher accuracy on the supervised-sparse block than on the unsupervised-sparse block, regardless of order. In addition, accuracy on all blocks was positively related to language ability. This relationship did not vary by block or order. For reaction time, we saw and interaction between block and order, but no effect of language ability. The unsupervised-sparse block was by far the most difficult block for all participants who received it. Thus, this interaction may reflect this block difference crossed with learning effects. Participants who received the unsupervised-sparse block second were perhaps more comfortable with the task overall than participants who received the unsupervised-sparse block first, which lead to shorter reaction times for those receiving unsupervised-sparse second.

exp1_ordereffects.png

Figure 1. Sensitivity (d') for each block completed by each group for all order effects. Colors indicate which block was encountered first by each group. Points indicate means with error bars reflecting standard error. Shaded portions represent the distribution of sensitivity values; wider portions indicate more subjects with that sensitivity value.

order_rt.png

Figure 2. Reaction time (s) for each block completed by each group for all order effects. Colors indicate which block was encountered first by each group. Points indicate means with error bars reflecting standard error. Shaded portions represent the distribution of reaction times; wider portions indicate more trials with that reaction time.

oe3_lang.png

Figure 3. Language ability is a significant predictor of sensitivity (*d'*) for blocks in the third order effect (all containing sparse stimuli).

1.4 Discussion

In this experiment, I tested three different order effects to see whether the order in which an individual engages the two category learning systems affects their learning using that system. The two manipulations in the statistical density task encouraged participants to use a particular system in two ways (learning type and stimulus type; see Table 2 for a summary). The analysis investigated whether block order affected performance in blocks where both the learning type and the stimulus type engaged the same system. The second analysis tested the effect of block order on performance when all stimuli were dense, and the third analysis did the same for only sparse stimuli.

1.4.1 Order Effects

The three analyses revealed what appears to be a general learning effect. It is most apparent in the first analysis, which showed that when both learning type and stimulus type engage the same category learning system, performance is better on the second block than on the first. Group 1 (unsupervised-dense first) showed slower reaction times in their first block, while Group 2 (supervised-sparse first) showed poorer accuracy in their first block, even though the first block for each of these groups was different. This result was also seen in a block by order interaction in the third analysis, where the difference between blocks in reaction times attenuated when the more difficult block (unsupervised-dense) was encountered second. Finally, while there were no significant effects in the second analysis, the mean reaction times are numerically higher for first blocks than for second.

The core hypothesis for this experiment was that engaging the hypothesis-testing system before the associative system would lead to reduced performance during associative blocks, but the reverse effect would not appear. This hypothesis was based on previous research that showed that when participants were required to switch between categories built using different category rules, they tended to rely more on executive function even if they were not actually switching between rule types (**?**). Other research also has shown that when participants are asked to learn a hybrid category that combines different rule types, they end up using only a simple rule-based strategy (**?**). Thus, when individuals are bombarded with cues towards different systems on a trial-to-trial basis, they default to the more explicit strategies, reflecting reliance on the hypothesis-testing system. However, this type of result was not found in the current study. Instead of defaulting to the hypothesis-testing system and thus showing reduced performance on unsupervised or dense blocks that occurred second, better performance was almost always seen in second blocks. This may reflect a broad learning effect that was not seen in prior studies.

Differences in experimental paradigm may at least partially explain why this study shows learning effects while other studies show reliance on a single system. In the studies mentioned above, stimulus characteristics encouraged participants to switch between systems on a trial-by-trial basis. In the current study, trials were blocked and a single system was engaged for that block. Participants had short transition periods between blocks were new instructions and examples or rules were presented. The results presented here suggest that these transition periods were sufficient for participants to switch to a new system as stimulus and task demands changed.

1.4.2 Individual Differences

The original hypothesis for this analysis was that individuals with poorer language ability would show stronger order effects than those with better language ability. My prior research showed that poor comprehenders (individuals with poor reading comprehension despite intact decoding ability) showed difficulty switching away from suboptimal learning strategies that were developed in the absence of guided instruction (**?**). Thus, at the very least I expected to see an interaction between language ability and order such that individuals with better language skills would show great recovery in performance when switching between unsupervised and supervised tasks, while those with poorer language skills would not show such a difference. However, no interactions between language ability and order were found.

The third analysis revealed the only significant effect of language ability. This analysis showed that language ability was positively related to accuracy when all items in both blocks were sparse. Recall that sparse items are best learned by the hypothesis-testing system. Since both blocks in the third order effect analysis were sparse, an individual could succeed by only using this system. Furthermore, language's effect on category learning is often found only for the hypothesis-testing system. Thus, this result is most in line with previous findings.

This finding is especially interesting because it is one of the first to relate category learning performance to individual differences in language ability in an adult sample. This topic has been much more extensively studied in children and infants. For example, vocabulary and categorization have been shown to be positively correlated in 20-month-olds (**?**) and 24-month-olds (**?**). In addition, infants' categorization ability at 12 months predicts their concurrent and future (18 month) vocabulary size (**?**). However, much of the individual differences category learning literature focuses on things like working memory or strategy use rather than language ability. Thus, this study is one of the first to find that individual differences in language ability are related to categorization accuracy for novel rule-based categories in adults.