1 Experiment 1

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 2. Unequal group sizes result from lost data due to technical errors.

Table 1. Group sizes for each order

Effect	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.

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.

Figure 1. Examples of learning type and category type manipulations for category learning experiment.

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 tested three main order effects. First, I tested order effects for the matching conditions (unsupervised-dense and supervised-sparse).The second order effect used unsupervised-dense and supervised-dense blocks. Finally, the third order effect tested the same sparse stimuli, testing unsupervised-dense and supervised-sparse blocks. This design led to six possible order groups that each participant could be placed into. See Table 2 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). 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 2. Block orders for statistical density task

Effect	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 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 3 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 simply 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 will use the recalling sentences subtest from the Clinical Evaluation of Language Fundamentals - Fourth Edition (CELF; **?**). This test covers 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.

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, we restricted participants to 10 minutes in this task. Since this administration is different than the standard administration, we do not use standard scores. Raw scores are calculated by counting the number of correct answers given within 10 minutes.

Table 3. 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

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.

1.3.1 Behavioral Measures

For basic descriptive statistics on the behavioral measures, seeTable **??**. First, I used 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. Transformation was done using the R package **caret**. The remaining measures (TOWRE, WA) were not skewed and thus were simply scaled and centered.

Next, I constructed a correlation matrix between all of the behavioral measures. All of the measures were significantly positively correlated, 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 $\chi^2(10)$ = 239.14, p <0.001. These measures suggest that the measures were suitable for a PCA. The first component in the PCA accounted for 47.86% of the variance and had an eigenvalue of 2.39. All of the factor loadings for these measures were quite similar, ranging from -0.41 to -0.51. The second factor accounted for an additional 20.6% of the variance and had an eigenvalue of 1.03. 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 available measures. No subject was missing more than 1 measure. This composite measure was then scaled but not centered.

Table 4. 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

1.3.2 Order Effect 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.

I used linear mixed-effects models to examine the effects of block and order on accuracy at test. The relationship between accuracy and block/order showed significant variance in intercepts across participants *SD* = 0.26. Adding block and order as fixed effects significantly increased model fit, $\chi^2(2)$ = 13.04, *p* = 0.001. Adding the interaction between block and order further improved model fit $\chi^2(1)$ = 6.05, *p* = 0.014. Thus, the final model had fixed effects of block, order, and the interaction between block and order as well as random intercepts for subject.

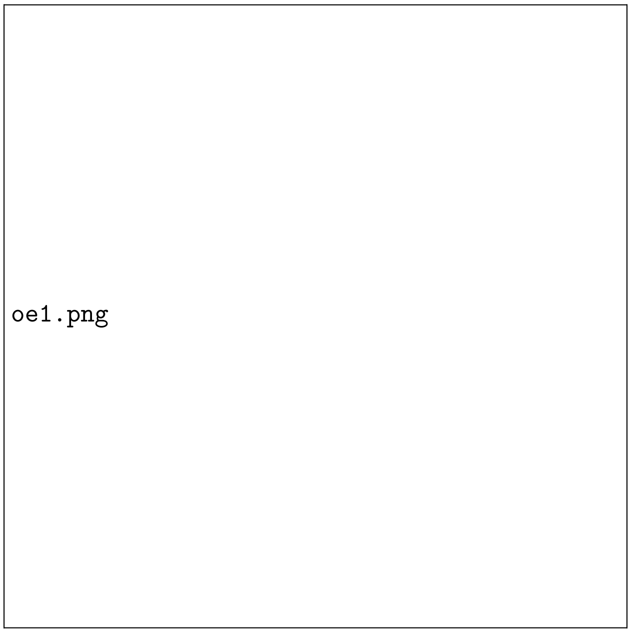

Figure 2. Accuracy (d') for each block completed by each group for the first order effect. Colors indicate which block was encountered first by each group. Points indicate means with error bars reflecting standard error. Shaded portions represent the distribution of accuracy values; wider portions indicate more subjects with that accuracy value.

The final model revealed three significant effects. First, there was not a significant main effect of block, *b* = -0.52, *SE* = 0.32, *t*(74) = -1.60, *p* = 0.12. This model also showed a significant main effect of order, *b* = -0.59, *SE* = 0.16, *t*(141) = -3.76, *p* = 0.0002. Finally, there was a significant interaction between block and order, *b* = 0.51, *SE* = 0.20, *t*(72) = 2.48, *p* = 0.016. This 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, *b* = -0.0064, *SE* = 0.11, *t*(34) = -0.057, *p* = 0.95. When the hypothesis testing system was used first (supervised-sparse first), there is a significant effect of block, *b* = 0.50, *SE* = 0.17, *t*(37) = 2.92, *p* = 0.0059. Inspection of means shows that when participants complete the supervised-sparse block first, performance on the supervised-sparse block is lower than in the unsupervised-dense block (see Figure 2).

1.3.3 Order Effect 2: Dense Stimuli

The second order effect analysis compared groups 3 and 4. All participants learned only dense categories, with the order of learning types differing between groups. Again, I used linear-mixed effects models to investigate the effects of block and order on accuracy at test. The variance in intercepts across participants had a standard deviation of 0.62. Adding the fixed effects to the model did not significantly improve fit $\chi^2(2)$ = 0.24, *p* = 0.89. Inspection of coefficients confirmed this finding. Block was not a significant predictor of accuracy, *b* = 0.03, *SE* = 0.10, *t*(145) = 0.27, *p* = 0.79. Similarly, order was not a significant predictor of accuracy, *b* = -0.04, *SE* = 0.10, *t*(145) = -0.11, *p* = 0.68. Thus, accuracy at test on dense categories was similar regardless of training type or block order (see Figure 3).

oe2.png

Figure 3. Accuracy (d') for each block completed by each group for the second order effect.

1.3.4 Order Effect 3: Sparse Stimuli

The third order effect investigated differences in learning sparse categories based on learning type order, using data from groups 5 and 6. I used the same type of linear mixed-effect models as the prior two order effects. Random intercepts for subject had a standard deviation of 0.11. Adding block as a fixed effect significantly increased model fit, $\chi^2(1)$ = 59.44, *p* $<$0.00001. Adding order to this model did not further improve model fit $\chi^2(1)$ = 0.05, *p* = 0.82. Thus, the final model had a fixed effects of block as well as random intercepts for subject, but no fixed effect of order or interaction between block and order. This model revealed a significant main effect of block, *b* = -1.14, *SE* = 0.13, *t*(72) = -8.91, *p* $<$0.00001. Inspection of means showed that participants exhibited better performance in supervised-sparse blocks than in unsupervised-dense blocks (see Figure 4).

1.3.5 Exploratory Order Analyses

An interesting feature of this experimental design is that both manipulations (learning type and category type) push individuals towards a certain category learning system. Supervised and sparse blocks encourage use of the hypothesis-testing system, while unsupervised and dense blocks evoke the associative system. Thus, mismatch blocks (i.e., unsupervised-sparse and supervised-dense) have conflicting information on which category learning system to use and thus likely are less effective at evoking that system. To investigate this possibility, I completed some exploratory analyses.

First, I compared unsupervised-dense blocks completed by groups 2 and 4. Group 2 completed a supervised-sparse (matching, hypothesis-testing) block before their supervised-dense block, while group 4 completed a supervised-dense (mismatch, hypothesis-testing) block before their unsupervised-dense block. If matching blocks more strongly evoke the category learning system and the hypothesized order effect (where activating the hypothesis-testing system first interferes with later use of the associative system) holds, then performance in group 2 on the unsupervised-dense block should be worse than performance in group 4 on the same block. A two-sample *t*-test indicated that this hypothesis did not hold – the two groups had equivalent performance (*t*(73) = -0.62, *p* = 0.54).

I extended this analysis by doing the sample thing for groups 1 and 5, who both completed the supervised-sparse block second. Group 1 completed an unsupervised-dense (match, associative) block before their supervised-sparse block and group 4 completed an unsupervised-sparse (mismatch, associative) block first. Again, a two-sample *t*-test indicated that the two groups had equivalent performance (*t*(69) = 0.36, *p* = 0.71).

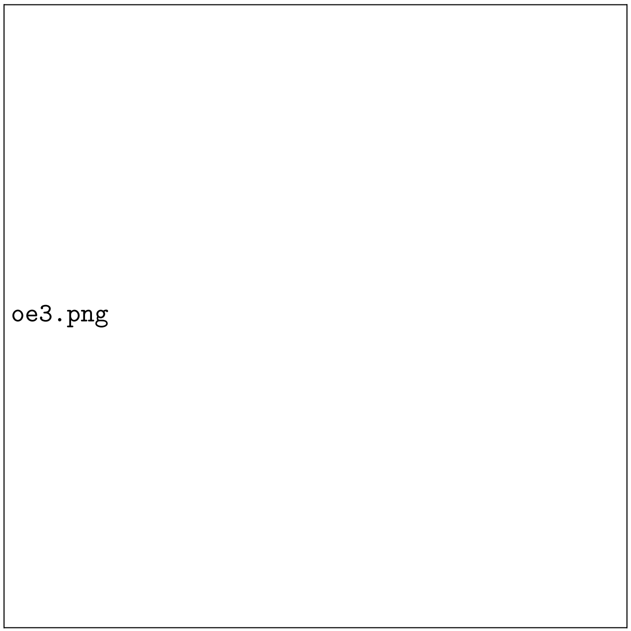

Figure 4. Accuracy (d') for each block completed by each group for the third order effect.

As an additional check, I looked at two more comparisons. First, I compared the unsupervised-dense blocks for the two groups who completed it first, group 1 and group 3. There should be no difference between these groups on this block, since it was the first block each group encountered. A two-sample *t*-test confirmed this hypothesis (*t*(70) = 0.097, *p* = 0.92). I then checked the same thing for the supervised-sparse blocks within groups 2 and 6. Interestingly, these two groups were found to be different (*t*(54) = -3.43, *p* = 0.001).

1.4 Discussion

1.4.1 Order Effects

The primary analyses looked at three different order effect. Each effect compared different ways to engage the hypothesis testing system before the associative system and vice versa. The first order effect used category learning blocks whose learning type and category type matched (i.e., supervised-sparse/hypothesis-testing, unsupervised-dense/associative). A significant interaction was found between block and order. Individuals who completed the unsupervised-dense/associative block first showed similar performance on both blocks. However, participants who completed the supervised-sparse/hypothesis-testing blocks first showed reduced performance in the supervised-sparse/hypothesis-testing block, with considerable recovery by the time they got to the unsupervised-dense/associative blocks. This result may be spurious, however, because the individuals in group 2 showed atypically low accuracy on their first block (supervised-sparse/hypothesis-testing), as compared to participants in group 6 that received the same block first and exhibited higher performance.

Overall, few order effects were found in accuracy. Most performance was close to ceiling. The main effect of block found in the third order analysis indicated that the unsupervised-sparse block may be more difficult overall. This is consistent with prior between-subjects research showing the worst performance in the unsupervised-sparse condition (**?**).

1.4.2 Individual Differences