Dissociable behavioural outcomes of visual statistical learning

Brett C. Bays, Nicholas B. Turk-Browne, and Aaron R. Seitz

Supplemental data

Reliability of learned and non-learned classifications

The pairs which were classified as learned and non-learned in the search task of Experiment 1 were classified by averaging the RTs of a pair's particular first-position shape across all six blocks of the search task and comparing that to the average RT of that pair's second-position shape across all six blocks of the search task. Each shape was tested once in each block. Although each trial of a block only included a single target, and thus no within-trial comparison of a pair's first- and second-position shapes is possible, it is still possible to compare the RTs of each pair's two shapes within a block. (This assumes that a subject's RT will be relatively stable over the block and that the randomized position of the target shape within the trial presentation will not drastically affect the RT.) By counting the number of instances in which a pair's second shape had a faster RT than its first shape on a block-by-block basis, we can obtain a crude measure of reliability of the learned vs. non-learned classifications. We would expect when comparing pairs classified as learned to pairs classified as non-learned that the learned pairs would comprise a higher number of blocks in which the second shape has a lower RT than the first shape. Figures S1A and S1B show the distribution of within-block item comparisons in which the RT for the second-position shape is lower than the RT for the first-position shape for the 106 pairs classified as learned and for the 79 pairs classified as non-learned in Experiment 1, respectively. It is possible for a given pair to have a second-position shape with a lower RT than its first-position shape up to six times - once in each block of the search task. The ordinate of both figures corresponds to the total count within each bin of the histograms. The mean number of blocks with a faster second-position RT than first-position RT for the learned pairs is 3.80 (Fig. S1A, dashed green line; SD = 1.046) and for the non-learned pairs is 2.18 (Fig. S1B, dashed green line; SD = 1.059); an independent-samples t test reveals a highly significant difference between these two means ($t(183)=10.39$, $p<0.001$, Cohen's $d = 1.54$).

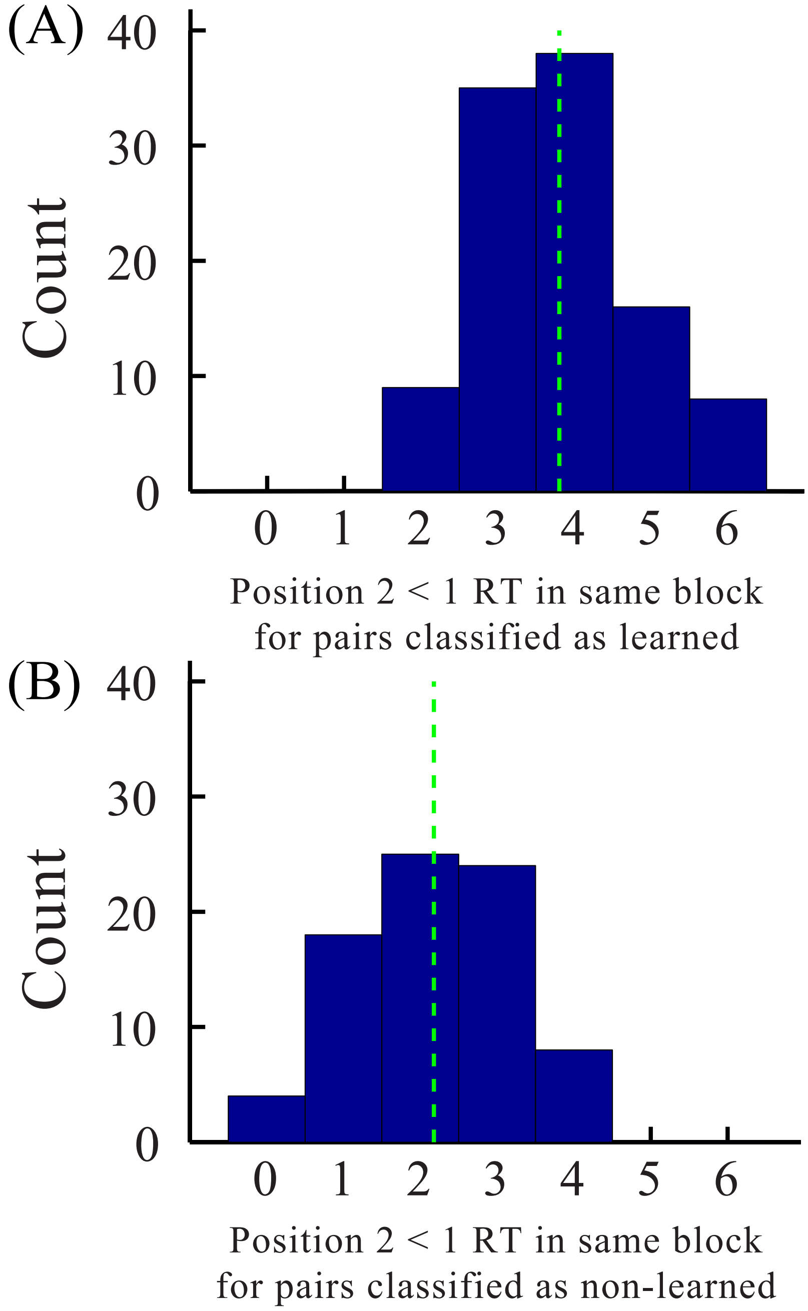

Figure S1 - Histograms displaying the number of times the RT to a second-position shape of a pair was lower than the RT to a first-position shape within the same block of the search task of Experiment 1 for (A) each of the 106 learned pairs and (B) each of the 79 non-learned pairs. The dashed green line on each figure represents the mean of the distribution, (A) 3.80 and (B) 2.18.

Statistical analyses

Given that subjects differed in the number of “learned” and “non-learned” pairs in the search task, we employed a modified 2x2 (position: first/second X learning status: learned/non-learned) factorial ANOVA approach to analyze these data, which included subtracting each participant's mean RT (across all shapes) from the RT for each shape. This method is statistically equivalent to a repeated-measures design (removing between-subject variance but retaining within-subject comparisons) while also permitting unequal cell counts (and was verified on balanced datasets). This approach was not an attempt to increase statistical power but rather was a direct method designed to address unequal cell counts, which themselves were an expected result of the coding procedure used. Analyzing data containing unequal cell counts is unadvisable (and sometimes impossible) with traditional repeated-measures designs and obtaining equal cell sizes in these experiments could only occur if each participant had the exact same number of pairs classified as learned and non-learned, regardless of how many total pairs were used.

Overall detection task results

Prior to splitting the pairs based on the post-test of each experiment, the detection task generally showed an overall pattern of lower RTs for second compared to first positions. However, the reliability of this effect was variable across experiments.

Experiment 1 (Figure S2) displayed a non-significant decrease in RT for second vs. first positions (666.7 vs. 674.5 ms, respectively; $t(36)=1.08$, $p=0.29$, Cohen's $d=0.18$), and a non-significant increase in accuracy for second vs. first positions (85.6 vs. 84.8%, respectively; $t(36)=0.69$, $p=0.50$, Cohen's $d=0.12$).

Experiment 2 (Figure S3) showed a trend for a decrease in RT for second vs. first positions (596.3 vs. 603.6 ms, respectively; $t(40)=1.70$, $p=0.097$, Cohen's $d=0.26$), however also a significant decrease in accuracy for second vs. first positions (88.3 vs. 90.7%, respectively; $t(40)=2.71$, $p=0.0098$, Cohen's $d=0.43$); this may represent a speed accuracy trade-off.

Experiment 3 (Figure S4) displayed a significant decrease in RT for second vs. first positions in the intact condition (550.0 vs. 557.2 ms, respectively; $t(55)=2.05$, $p=0.046$, Cohen's $d=0.27$) and a non-significant increase in accuracy for second vs. first positions in the intact condition (88.7 vs. 88.3%, respectively; $t(55)=0.44$, $p=0.66$, Cohen's $d=0.071$). There was no significant difference between second vs. first positions in the foil condition for RT (561.0 vs. 560.7, respectively; $t(55)=0.076$, $p=0.94$, Cohen's $d=0.012$) or accuracy (89.4 vs. 89.7%; $t(55)=0.34$, $p=0.73$, Cohen's $d=0.058$). Mismatched second shapes displayed a small but significant increase in accuracy compared to intact first shapes (89.9 vs. 88.3%, respectively; $t(55)=2.35$, $p=0.023$, Cohen's $d=0.32$), but no significant difference in RT (559.1 vs. 557.2 ms, respectively; $t(55)=0.46$, $p=0.64$, Cohen's $d=0.064$).

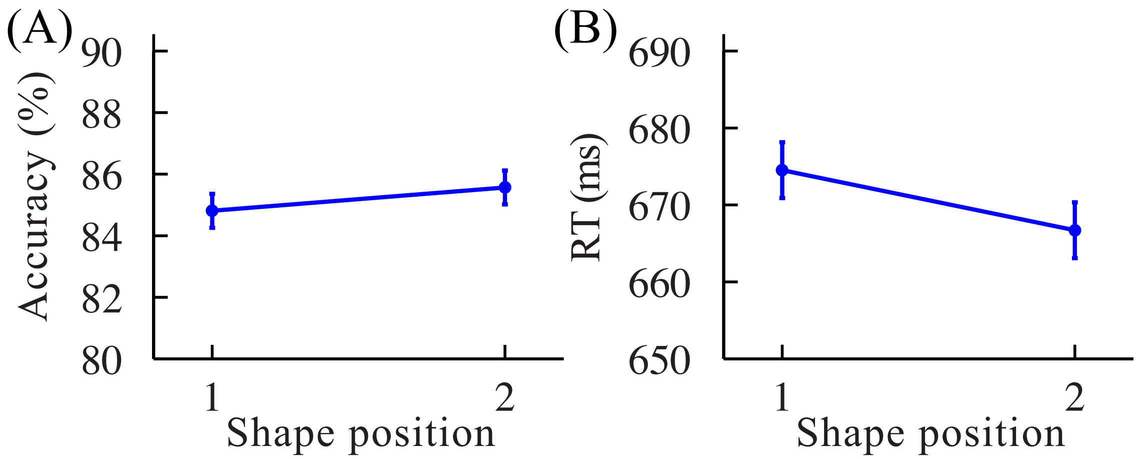

Figure S2 - Detection task results in Experiment 1, in terms of (A) accuracy and (B) RT. Error bars reflect +/-1 within-subjects SEM (Loftus & Masson, 1994). N = 37.

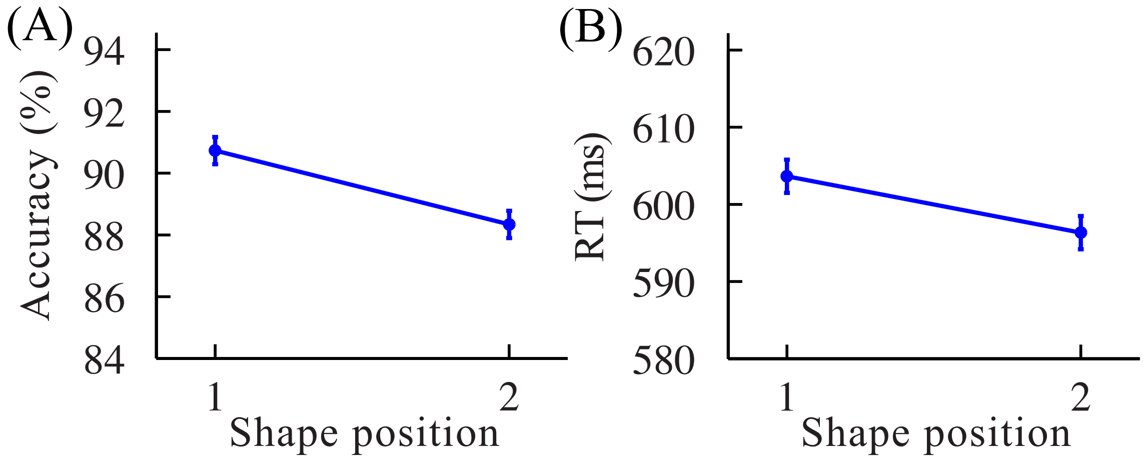

Figure S3 - Detection task results in Experiment 2, in terms of (A) accuracy and (B) RT. Error bars reflect +/-1 within-subjects SEM. N = 41.

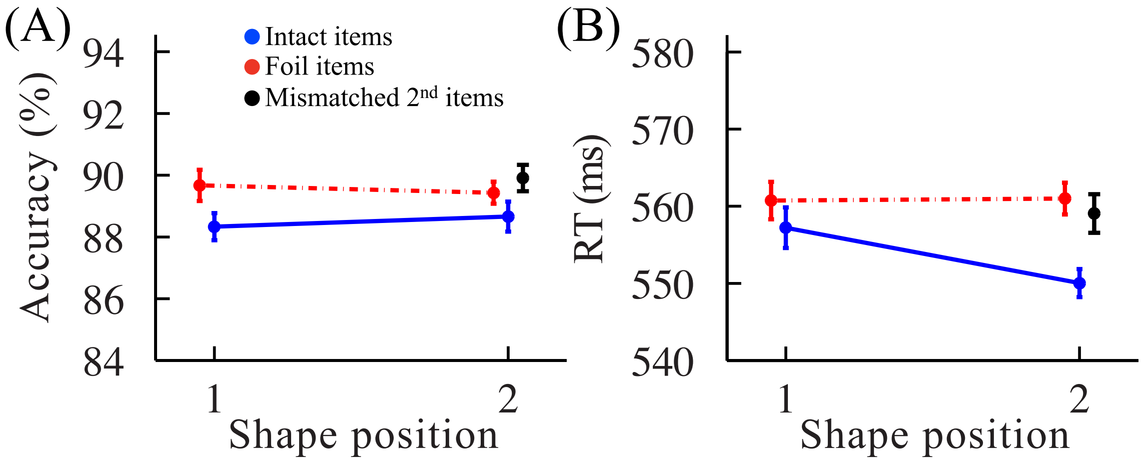

Figure S4 - Detection task results in Experiment 3. "Intact items" were items presented in their associated pairs (solid blue line). "Foil items" were items presented with either the first or second item replaced with a foil unseen in the first ten blocks of exposure (dashed red line). "Mismatched 2nd items" were items in which the second item of a pair was replaced with the second item of another pair (black point). Error bars reflect +/- 1 within-subject SEM. N = 56.

Results using the full triplet structure

Although we do not believe that the full triplet structure was learned in our experiments, it is nevertheless informative to perform the analyses for Experiments 1 and 2 using the full triplet structures of the experiments. Figure S5 shows the search task results for Experiment 1, in which there is an expected monotonic decrease in RT across item position (535ms vs. 520ms vs. 502ms, respectively; $F(2,72)=9.42$, $p=0.0002$, $\eta^2=0.021$). In order to classify triplets as "learned" or "non-learned", a metric similar to that used with pairs was applied, in which a triplet was classified as "learned" if it showed a monotonic decrease in RT across item position (i.e., item 1 RT > item 2 RT > item 3 RT) and "non-learned" if the triplet did not meet this standard. Although this rule is similar to the rule used for pairs, it results in different assignments to triplets than would have been made via a decision based on the first two items of that triplet. This can be seen in Figure S6, which shows accuracy and RT for the detection task of Experiment 1 for triplets as defined by whether they were learned or not in the search task. Although the trend is similar to the patterns seen from analyzing pairs, due to the different categorizations there is not as clear of a separation between the items classified as learned and non-learned in accuracy (interaction $F(2,549)=0.71$, $p=0.49$, $\eta^2=0.0026$) or RT (interaction $F(2,549)=2.86$, $p=0.058$, $\eta^2=0.01$).

Experiment 2 utilized a recognition task to measure learning after the detection task and, as reported in the main text, 139 triplets were correctly identified and 66 triplets were not correctly identified. These classifications of the triplets are the same as the classifications for the pairs reported in the main text since the recognition judgment was made using triplets and does not change whether we analyze the full triplet or just the first two items of that triplet in the detection task. The results for the full triplets in the detection task, as split by whether they were classified as learned or non-learned based on their identification in the recognition task, and split by whether they were judged as “Familiar” or “Remember” (see Experiment 2 Methods in main text), can be seen in Figure S7. Patterns similar to those seen in the pair analyses emerge, with a significant interaction for RT (interaction $F(4,606)=2.56$, $p=0.038$, $\eta^2=0.016$) but not for accuracy (interaction $F(4,606)=1.31$, $p=0.26$, $\eta^2=0.0084$). Collapsing the Familiar and Remember judgments (as in Figure 7 of the main text) reveals similar effects, with a significant interaction for RT (interaction $F(2,609)=4.74$, $p=0.009$, $\eta^2=0.015$) but not for accuracy (interaction $F(2,609)=1.7$, $p=0.18$, $\eta^2=0.0055$).

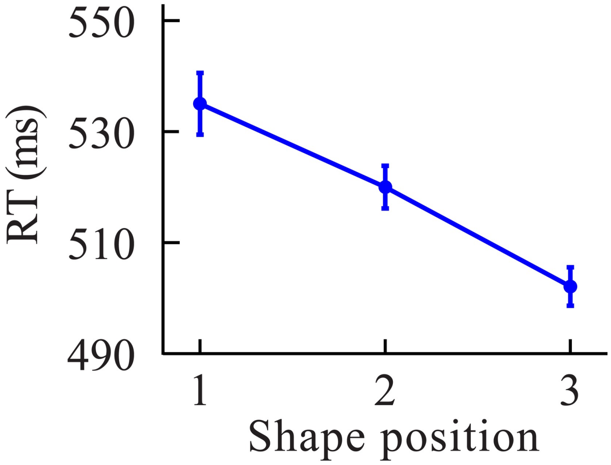

Figure S5 - Mean RT for the full triplet structure in the search task of Experiment 1. Error bars reflect +/- 1 within-subjects SEM. N = 37.

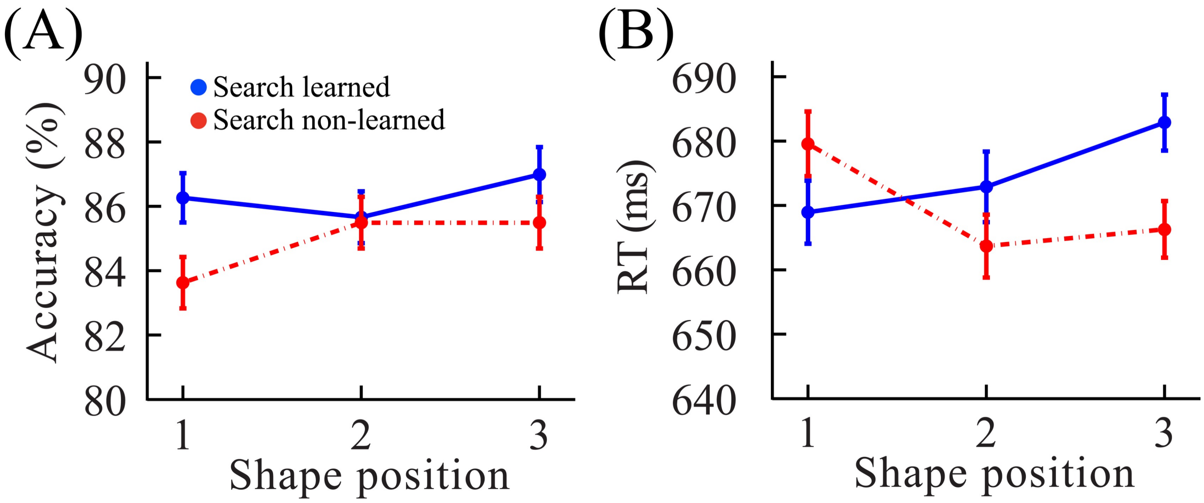

Figure S6 - Detection task results for the full triplet structure in Experiment 1 split by the search task, in terms of (A) accuracy and (B) RT. "Search learned" were triplets that demonstrated learning in the subsequent search task (solid blue lines). "Search non-learned" were triplets that did not demonstrate learning in the subsequent search task (dashed red lines). Error bars reflect +/- 1 within-subjects SEM. N = 37. (N of triplets in blue curves = 83; N of triplets in red curves = 102.)

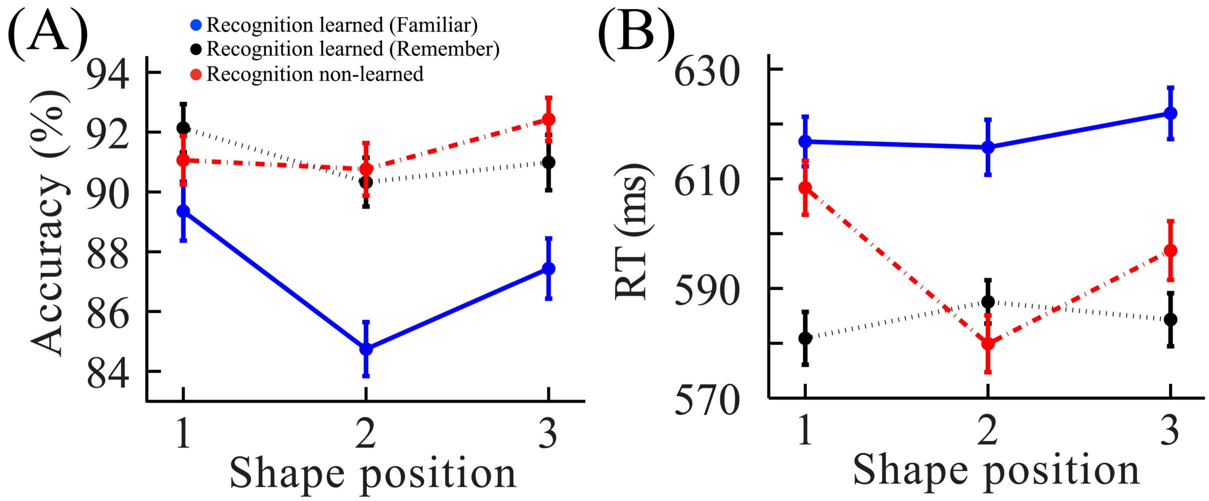

Figure S7 - Detection task results for the full triplet structure in Experiment 2 split by the Familiar/Remember ratings in the recognition task, in terms of (A) accuracy and (B) RT. "Recognition learned (Familiar)" were triplets correctly identified in the subsequent recognition task and given a "Familiar" rating (solid blue lines). "Recognition learned (Remember)" were triplets correctly identified in the subsequent recognition task and given a "Remember" rating (dotted black lines). "Recognition non-learned" were triplets not correctly identified in the subsequent recognition task (dashed red lines). Error bars reflect +/- 1 within-subjects SEM. N = 41. (N of triplets in blue curves = 78; N of triplets in black curves = 61; N of triplets in red curves = 66.)