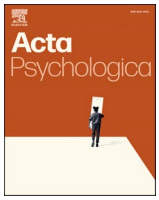
Acta Psychologica 244 (2024) 104187

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy



Release from response interference in color-word contingency learning

Brady R.T. Roberts *, Noah D. Forrin, David McLean, Colin M. MacLeod

Department of Psychology, University of Waterloo, Waterloo, Canada

ARTICLE INFO

Keywords:
Learning
Contingency learning
Response interference
PEP 2.0 model

ABSTRACT

In identifying the print colors of words when some combinations of color and word occur more frequently than others, people quickly show evidence of learning these associations. This *contingency learning effect* is evident in faster and more accurate responses to high-contingency combinations than to low-contingency combinations. Across four experiments, we systematically varied the number of response-irrelevant word stimuli connected to response-relevant colors. In each experiment, one group experienced the typical contingency learning paradigm with three colors linked to three words; other groups saw more words (six or twelve) linked to the same three colors. All four experiments disconfirmed a central prediction derived from the Parallel Episodic Processing (PEP 2.0) model (Schmidt et al., 2016)—that the magnitude of the contingency learning effect should remain stable as more words are added to the response-irrelevant dimension, as long as the color-word contingency ratios are maintained. Responses to high-contingency items did slow down numerically as the number of words increased between groups, consistent with the prediction from PEP 2.0, but these changes were unreliable. Inconsistent with PEP 2.0, however, overall response time did not slow down and responses to low-contingency items actually sped up as the number of words increased across groups. These findings suggest that the PEP 2.0 model should be modified to incorporate response interference caused by high-probability associations when responding to low-probability combinations.

Associative learning may well be the cornerstone of all learning. Connecting two entities, whether two stimuli, two responses, a stimulus and a response, or a variety of other combinations (e.g., a stimulus and its context, two dimensions of a single stimulus or event) routinely underpins more extensive learning. Such learning is at the core of both classical and operant conditioning and has been investigated in many animal species. In their review, Le Pelley et al. (2016) argued that, in human associative learning, attention is biased toward highly predictive stimuli even when their processing is considered to be automatic, and that this bias is greater the more predictive a stimulus is of some outcome (also see Mackintosh, 1975).

One situation where associative learning is readily studied in humans is contingency learning—learning from a statistical correlation between two entities that each predicts the other. Such a correlation can speed responding, make responding more accurate, and even influence stimulus evaluation and judgments of causality. It can also happen extremely quickly (Lewicki, 1985), suggestive of 'preparedness' to learn such connections (Seligman, 1970). Often, consistent with the idea of prediction as expectation, human contingency learning has been investigated in sequential circumstances, where a prior stimulus predicts a subsequent stimulus (see, e.g., Shanks, 2007). But it can also be studied in simultaneous circumstances where one stimulus—or one dimension of a stimulus—is correlated with another co-occurring stimulus or dimension. Here, one of the stimuli or dimensions is response-relevant whereas the other is response-irrelevant, allowing for investigation of contingency influences that can be less obvious to the participant than is often the case with successive relations.

A paradigm that provides a very simple way to investigate human contingency learning is the color-word contingency paradigm introduced by Schmidt et al. (2007). In this situation, a combined color-word stimulus is presented on each trial, with the color dictating the response and the word being response-irrelevant. That is, in the typical color-word contingency paradigm, participants are told to respond as quickly and as accurately as they can to the print color of each word. Typically, only three colors and three words are used (e.g., the colors red, yellow, and green, and the words *mouth*, *under*, and *plate*). The contingency manipulation resides in the assignment of each word preferentially to one of the colors such that, for example, *mouth* appears 80% of the time in red and only 10 % of the time in each of the other two colors. In this example, *mouth* in red is a high-contingency item and

* Corresponding author at: Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada.
E-mail address: bradyrtroberts@gmail.com (B.R.T. Roberts).

https://doi.org/10.1016/j.actpsy.2024.104187
Received 20 October 2022; Received in revised form 5 February 2024; Accepted 12 February 2024
Available online 16 February 2024

mouth in yellow or in green is a low-contingency item. Here, we will refer to these respectively as the HI and LO conditions. Very quickly, participants show faster (and more accurate) responding on HI trials than on LO trials, and this is true whether a participant is or is not aware of the contingency, although the difference is enhanced by awareness (Schmidt, De Houwer, 2012a, 2012b). The effect is robust and stable, and there is now a considerable literature exploring this simple situation (for reviews, see MacLeod, 2019; Schmidt, 2021).

In 2013, Schmidt introduced a model intended to capture this type of simple learning. His initial Parallel Episodic Processing (PEP) model has since been updated to PEP 2.0 by Schmidt et al. (2016; see also Schmidt, 2018). The model assumes that as individual trials occur in an experiment, they are stored in episodic memory and then are routinely retrieved on subsequent trials, influencing how those subsequent trials are handled. PEP 2.0 is a thus member of the class of instance theories of learning (e.g., Logan, 1988), although in PEP 2.0 instances are recovered simultaneously rather than there being a 'race' among instances. On each trial—with a trial stored as an episode—both the relevant (e.g., the font color) and the irrelevant (e.g., the word) elements are encoded, along with the response made (e.g., pressing a particular response key). On any given trial, more recently experienced instances exert more influence on processing than do less recently experienced instances.

Applied to the color-word contingency learning paradigm, PEP 2.0 handles the basic contingency learning effect by predicting a benefit for the HI trials. Using our example, because *mouth* has most often been seen in red previously (i.e., 80% of the time), when the current trial contains *mouth*, there will be more retrieved instances in memory that point to *mouth* in red than to *mouth* in another color. If indeed the current trial is *mouth* in its high-contingency color (red), PEP 2.0 predicts that the red response will be facilitated due to many past episodes biasing the response decision in favor of the high-contingency color. If the current trial is *mouth* in any other (low-contingency) color, PEP 2.0 suggests that the scarcity of matching instances in memory will lead to a weaker response bias toward those colors, resulting in a slower response decision than would occur on a HI trial. The result is a positive LO - HI difference both in latency and in accuracy of responses that Schmidt et al. (2016, p. 84) refer to as the "contingency learning benefit." More neutrally, we call this difference the *contingency learning effect*.

1. Costs for LOs in contingency learning

In the same year that their PEP 2.0 model became available, Schmidt and De Houwer (2016) also published an article that considered evidence of four main accounts to explain the contingency learning effect: (1) prediction benefit, (2) misprediction cost, (3) bidirectional cost, and (4) pure proportion. The prediction benefit account posits that a participant's expectations would ready a particular color response when a high contingency word was presented, leading to facilitation on HI trials. The misprediction cost account says that not only will HI trials be facilitated, as in the prediction benefit account, but that there will also be a *cost* to inaccurately predicting a response given a low-contingency word (that is, responses on LO trials will be slowed due to having to overcome the incorrect prepared response). The bidirectional account is largely the same as the misprediction account except that it additionally assumes that colors can be predictive of words as well, and therefore when presented with a color that is usually highly predictive of a different word, responses to the low-contingency words will be slowed. Finally, the pure proportion account suggests that all HI and LO responses are facilitated to some degree, but the extent to which this occurs depends on the proportions of these trials in what has gone before (therefore, HI trials are more facilitated due to their higher prominence in the set).

Critically, of these four accounts, only the middle two (misprediction cost and bidirectional cost) posit a *cost* for LO trials. The other two accounts (prediction benefit and pure proportion) cast the contingency learning benefit as entirely facilitative. Schmidt's PEP 2.0 account aligns with the pure proportion account, suggesting that there is no cost to LO trials, only less facilitation. Yet several studies have reported costs to LO trials when compared with neutral baselines that have no contingency, including two examples from the MacLeod laboratory (Forrin & MacLeod, 2018; Lin & MacLeod, 2018) and one from the Schmidt laboratory (Schmidt & De Houwer, 2016, Experiment 1). In a 2021 review, however, Schmidt argues that investigations of medium contingency trials have revealed that participants do *not* exhibit response biases dependent on response expectations. The field is therefore undecided as to whether costs for LO trials are due to *response interference* (or '*response competition*', as the misprediction and bidirectional cost accounts would predict), or if instead there is *retrieval interference* (as Schmidt & De Houwer, 2016 refer to it) driving less facilitation for LO trials owing to their scarcity (as the pure proportion account and PEP 2.0 model would predict). Here, we aim to resolve this debate by implementing a variant of the color-word learning paradigm.

2. The present investigation

Previous studies of contingency learning have ordinarily been restricted to a single word having a high contingency association to each color; typically, each of three words to one of three colors. In our example, only *mouth* was presented 80% of the time in red, and each of the other two words were presented 80% of the time in only one of the two remaining colors (e.g., *under* in yellow and *plate* in green). But what would happen if more response-irrelevant words were added to the set without changing contingency proportions? That is, what would occur if there was a second word that was *also* highly contingent with the color red?

In the present experiment, instead of changing the contingency ratio as others have (e.g., Forrin & MacLeod, 2018), we introduced the manipulation of more than one word 'sharing' a given color, each of the shared words having high contingency to that color. Importantly, the manipulation used here added more HI trials *without* altering the proportion of LO trials (unlike what happens when the contingency ratio itself is changed).

A straightforward prediction from PEP 2.0 and the pure proportion account is that the contingency learning effect should remain the same because the proportion of instances in memory would still favor HI trials, leading to facilitation for HIs and little to no facilitation for LOs. Of course, the addition of more trials would mean that *all* responses should receive less facilitation when compared to the standard 3-word, 3-colors paradigm. In other words, there should be slower responses overall, but HI trials should still be faster than LO trials on average, leading to a contingency learning effect of consistent magnitude. A related account to the pure proportion account that is concerned instead with trial frequency (the pure frequency account; Schmidt & De Houwer, 2016) suggests that HI trials are facilitated not because they represent a higher proportion of trials but because there are more of them in general (referring to their raw trial count). This pure frequency account also predicts that the contingency learning effect should remain the same size as both HI and LO trials become scarcer in memory.

In our first experiment, we compared a 6-word condition to the typical 3-word condition. In subsequent experiments, we also included a 12-word condition to provide an extended test of the prediction. The purpose of this study was to directly test whether Schmidt's PEP 2.0 model and the pure proportion/pure frequency accounts can predict contingency learning beyond the conventional three words, three colors paradigm. We predicted that the contingency learning effect should be stable as word set size increases because both HI and LO instances matching the current trial should become equally scarce. In other words, the size of the contingency learning effect should not change when more response-irrelevant words are added to the study set if the typical 8:1:1 (HI:LO:LO) contingency ratio is maintained for each word, but responses on both HI and LO trials should become slower overall as matches to prior experiences become rarer.

3. Experiment 1

In this first experiment, we manipulated the number of words highly contingent with any given color—from the standard one word per color (3 words total) to two words per color (6 words total). Under PEP 2.0—and the pure proportion/pure frequency accounts more generally—the contingency learning effect was expected to remain the same size but overall response times were expected to slow.

3.1. Method

3.1.1. Participants

We collected a sample size that roughly matched or exceeded that of previous contingency learning studies.[1] A power sensitivity analysis is presented below in the Outlier Removal section. During the first academic term of 2018, 100 University of Waterloo undergraduate students took part in a single session in exchange for course credit. All had self-reported normal or corrected-to-normal vision. After data trimming, the final sample of 98 participants used in our statistical analyses was 83% female, with age ranging from 18 to 32 ($M = 20.04$, $SD = 2.22$). Due to missing or incomplete demographic data for some participants, these two figures are based on 77 and 90 participants, respectively.

3.1.2. Apparatus

The task was programmed using *E*-Prime 3.0 software (E-Prime, 2016). Stimulus presentation was controlled by a Windows-based computer with a 24″ color monitor set to 1920 × 1080 resolution. Responses were collected using a standard QWERTY keyboard.

3.1.3. Materials and design

The master list of six English words (*mouth, under, plate, bench, clock,* and *dream*) was made up of common high-frequency words (SUBTLEX-US Zipf scores ≥4; Brysbaert et al., 2019), each containing five letters. This list was randomized anew for each participant, with either a random 3 or all 6 words used, depending on condition. Word-to-color contingencies were also randomized across participants.

Using a 2 × 2 mixed design, the within-subject factor was contingency (HI vs. LO) and the between-subjects factor was number of words (3 vs. 6). Participants were randomly assigned to one of two between-subjects conditions, resulting in 50 participants per group. Participants in each group completed 600 trials, with each word appearing in one color 80% of the time (high contingency), and in each of the other two colors 10% of the time (low contingency; see Fig. 1). The only difference between the groups was that in the 6-word condition two words had a high-contingency connection to each color whereas in the 3-word condition only one word had a high-contingency connection to each color.

3.1.4. Procedure

Upon arrival (and after informed consent), participants saw on-screen instructions for the task and were asked after reading these instructions whether any clarification was required before proceeding directly to the trials. They were informed that on each trial a word would appear in the center of the screen in one of three colors, and that they were to press the key on the keyboard that corresponded to the color as quickly and accurately as possible. The keys—J = red, K = yellow, and L = green—had corresponding solid-colored stickers on them, resulting in key-to-color assignment being constant across participants.

After receiving a standard set of instructions read aloud by the experimenter, trials began immediately. Each trial began with a fixation cross (+) for 150 ms at the center of the screen, followed by a 150-ms blank screen. Then a word was presented in color in the center of the screen for a maximum of 2000 ms or until a response key was pressed. Stimulus words were presented in 18-pt Consolas lowercase font in one of the three colors (red, yellow, or green). Correct responses were followed immediately by the next trial; incorrect responses or 2000-ms timeouts led to a feedback screen displaying "XXX" for 500 ms before the next trial (as is standard procedure in this literature; see Schmidt et al., 2007, 2018). On all screens, a black background was used; instructions were presented in white font.

After all 600 trials were completed, and as has been standard practice in the color-word contingency learning literature (e.g., Schmidt et al., 2007; Schmidt & De Houwer, 2012c), participants were asked two questions to determine the extent of their awareness of the color-word relations. The first question was subjective and inquired about their awareness of the color-word congruency: "In this experiment, each word was presented most often in a specific color. Specifically, each word was presented mostly in either red, yellow, or green. Did you notice these relations? [Response options of YES or NO]."[2] The second question was objective: "In what color was the word [stimulus word] usually presented?" Each of the stimulus words was then presented one at a time in white font in a random order and the participant was to identify the high-contingency color associated with the word by pressing the corresponding color key. Following these questions, participants were debriefed verbally and with a written letter.

3.1.5. Outlier removal

To ensure accurate representation of average responses times and error rates, we subjected the data set to five cleaning steps that stem from a combination of procedures often used in prior research on contingency learning (Geukes et al., 2019; Schmidt, 2016; Schmidt et al., 2007, 2010) and in other RT-based work (e.g., Besner, McLean, Young, 2021a, 2021b). First, participants with corrupt data or ≥ 20% of trial missing were removed. Second, if a participant's overall error rate was ≥20%, that file was removed. Third, we removed any trials that were either anticipations (i.e., < 200 ms) or timeouts (i.e., exceeded the 2000-ms response deadline). [Afterward, we removed any participant who now had <80 % of trials remaining.] Fourth, only for correct responses, we removed any trials that were statistical outliers based on response time (i.e., ± 3 *SD*s away from the mean), as calculated separately for HI trials and LO trials within each participant. [Then, we once again removed any participant who now had <80% of trials remaining.] Fifth, we excluded participants whose mean response time or error rate (collapsed across HI and LO trials) was ±3 *SD*s away from the overall mean response time or error rate in their group. Each data cleaning step was performed using *R*; the syntax used for this process can be found on OSF (see the Transparency and Openness section below). Appendix Table 1 shows the small amount of data trimmed as a result of each data cleaning step.

In Experiment 1, we retained data from 98 of the 100 participants in our statistical analyses. A power sensitivity analysis using the *jpower* module (v. 0.1.2; Morey & Selker, 2020) for *jamovi* (v. 2.2.5; Şahin & Aybek, 2019) indicated that, with our final sample size of $N = 98$, we were powered to detect between-subjects effect sizes in independent samples *t*-tests as small as $d = 0.57$ ($f = .29$, $\eta_p^2 = .08$) and within-subject effect sizes in paired-samples *t*-tests as small as $d = 0.40$ ($f = .20$, $\eta_p^2 = .04$), both with 80% power ($\alpha = .05$, two-tailed).

[1] In each experiment, we used convenience sampling of undergraduate participants enrolled in psychology courses. As a result, our participants tended to be young, college educated women. This population matches that of much of the previous work on contingency learning, allowing for closer comparison to previous findings. Although we had no a priori reasons to believe that culture, language, race, or sex would affect the low-level cognitive processes thought to underlie contingency learning, the homogeneity of our sample necessarily limits the generalizability of any conclusions.

[2] Due to a programming error, responses to the subjective awareness question were not recorded in Experiments 1 and 3.

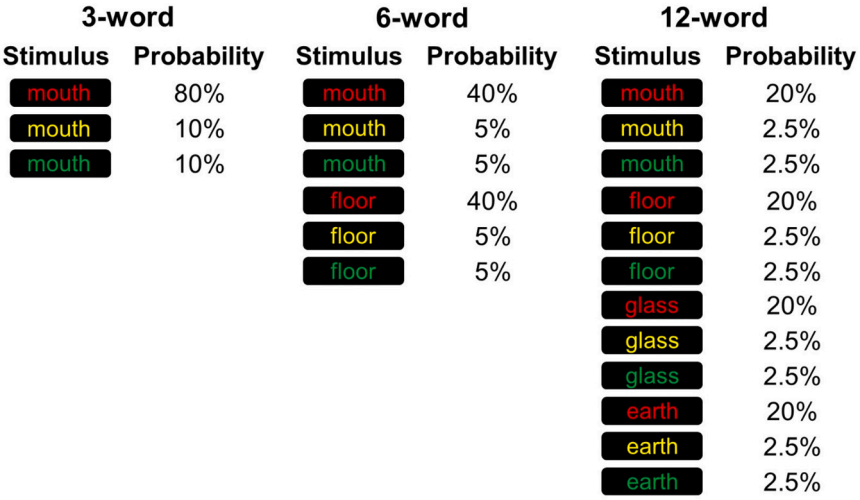


Fig. 1. Distribution of words connected to a single color (in this example, red) for each condition.
Note. The ratio of HI:LO:LO for each word is 8:1:1 across all color-word combinations. In Experiments 2–4, words were presented in uppercase.

3.1.6. Transparency and openness

The procedures and materials for this study were approved by the Office of Research Ethics at the University of Waterloo (projects #30188 & #41523). Although this study was not pre-registered, all program files, data, and statistical analysis code are posted on the Open Science Framework (OSF; https://osf.io/4x9v8/). Data cleaning and statistical analyses were performed using *R* (v. 4.1.1; R Core Team, 2020), enlisting the *afex* (v. 1.1–1; Singmann et al., 2022), *emmeans* (v. 1.8.1–1; Lenth et al., 2022), and *trend* (v. 1.1.4; Pohlert, 2020) packages. We report our rationale for sample size determinations, data exclusion steps and results, and all manipulations and measures in the study (in accordance with JARS; Kazak, 2018).

3.2. Results

For each experiment, a table presents descriptive statistics for both response time and error rate. In general, the error rate results were consistent with the response time results: Faster response times were associated with lower error rates, indicating no speed-accuracy trade-off. We focus on the response time inferential statistics in the main text, reporting the corresponding error rate inferential statistics in the Appendix. We also include in the Appendix the results concerning participants' subjective and objective awareness of the color-word contingencies.

Table 1
Experiment 1: response time and error rate.

Group	HI	LO	CL effect
Response times			
3-word	543 (60)	586 (78)	43 (39)
6-word	549 (72)	580 (80)	31 (26)
Error rates			
3-word	.034 (.023)	.062 (.046)	.029 (.036)
6-word	.038 (.024)	.059 (.053)	.021 (.045)

Note. This table reports mean response time (in ms) and mean error rate (with standard deviations for each in parentheses) for high-contingency (HI) and low-contingency (LO) trials, and the mean contingency-learning (CL) effect (low contingency – high contingency) in the two experimental groups.

The top half of Table 1 displays mean correct response times for HI and LO trials separately for the 3-word and 6-word groups. Before analyzing response time data, all error trials were removed (see Appendix Table 1 for error analyses). A 2 × 2 mixed analysis of variance (ANOVA)[3] was conducted with Contingency (HI vs. LO) as the within-subject factor, Group (3-word vs. 6-word) as the between-subjects factor, and response time (in ms) as the dependent measure. As expected, there was a significant main effect of Contingency, $F(1, 96) = 124.12$, $p < .001$, $\eta_p^2 = .56$, $BF_{10} > 100$, indicative of a robust overall contingency learning benefit. The main effect of Group was non-significant, $F(1, 96) < 0.01$, $p = .982$, $\eta_p^2 < .01$, $BF_{01} = 2.61$, with the Bayesian evidence for the null model being only anecdotal. The Contingency x Group interaction was non-significant, $F(1, 96) = 3.47$, $p = .066$, $\eta_p^2 = .04$, $BF_{01} = 1.03$. Consistent with our prediction, the contingency learning effect although numerically larger for the 3-word group (43 ms) than for the 6-word group (31 ms) was not significantly so. Because of our a priori hypotheses, we then conducted planned comparisons confirming significant contingency learning effects both in the 3-word group, $t(96) = 9.10$, $p < .001$, $d = 1.12$, $BF_{10} > 100$, and in the 6-word group, $t(96) = 6.63$, $p < .001$, $d = 1.17$, $BF_{10} > 100$ (see Fig. 2).

Next, because we also reasoned a priori that response times on HI and LO trials might differ between groups, we broke down the Contingency x Group interaction. Planned comparisons revealed that responses to HIs were non-significantly different between groups, $t(96) = 0.49$, $p = .628$,

[3] Throughout this article, Bayes factors were calculated using the *BayesFactor* (Morey & Selker, 2020) package for *R*, enlisting a default Jeffreys-Zellner-Siow (JZS) prior with a Cauchy distribution (center = 0, $r = .707$). This package compares the fit of various linear models. In the present case, Bayes factors for the alternative (BF_{10}) are in comparison to intercept-only models containing subject-level error, or to models containing subject-level error and both main effect terms in the case of Bayesian analyses of 2 × 2 interactions. Bayes factor interpretations follow the conventions of Lee and Wagenmakers (2013). Bayes factors in favor of the alternative (BF_{10}) or null (BF_{01}) models are presented in accordance with each preceding report of NHST analyses (i.e., based on a $p < .05$ criterion).

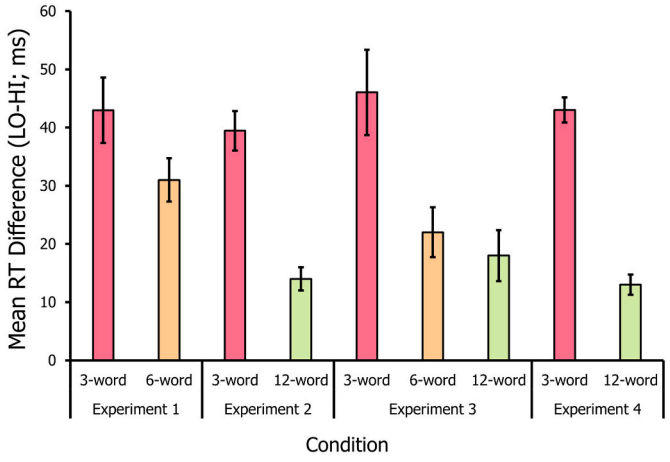


Fig. 2. Magnitude of the contingency learning effect in all four experiments. *Note.* Error bars = ± 1 *SE*.

$d = 0.10$, $BF_{01} = 4.23$, and that this was also true for LOs, $t(96) = 0.37$, $p = .713$, $d = 0.08$, $BF_{01} = 4.42$, with moderate Bayesian evidence for the null in each case.

3.3. Discussion

There are three principal results of interest. First, we replicated the basic contingency learning effect in the 3-word group—the version used in most prior studies. Second, despite the effect still being present in the 6-word group and not significantly different from that in the 3-word group, the effect size was numerically attenuated, which is counter to expectations from the pure proportion/pure frequency accounts. Third, and again inconsistent with these two accounts, we observed no overall slowing in moving from 3 to 6 words.[4]

In sum, our findings are inconsistent with predictions from pure proportion/pure frequency accounts, including the PEP 2.0 model. There was a numerical reduction in contingency learning effect size but with no overall slowing observed. Moreover, the LO trials seemed to speed up, a pattern not predicted by the pure proportion/pure frequency accounts. The following experiments explored these patterns further.

4. Experiment 2

The principal goal of Experiment 2 was to conceptually replicate and extend this 'shared associations' manipulation to further investigate whether contingency learning effect size would be altered as the number of stimuli on the response-irrelevant dimension increased. Therefore, there were again two groups, this time differing in whether they were presented with 3 words or 12 words in total. Based on the pure proportion/pure frequency accounts, we again expected general slowing of both HI and LO responses and a constant contingency learning effect.

4.1. Method

4.1.1. Participants

We again collected a sample size that roughly matched or exceeded that of previous contingency learning studies. A power sensitivity analysis is presented below in the Outlier Removal section. Our data collection stopping rule was based on the end of the academic term. During the final academic term of 2019, 166 University of Waterloo undergraduate students took part in a single session in exchange for course credit. All participants had self-reported normal or corrected-to-normal vision. After data trimming, the final participant sample of 163 participants used in our statistical analyses was 82% female, with age ranging from 17 to 27 (M = 19.55, SD = 1.81). Due to missing or incomplete data for four participants, the latter demographic figure is based on 159 participants.

4.1.2. Apparatus

The apparatus was identical to that in Experiment 1.

4.1.3. Materials and design

Participants were randomly assigned to condition, with 85 in the 3-word group and 81 in the 12-word group. The overall design followed that of Experiment 1 except that the word set was replaced by 12 new words. These new words were selected from the MRC psycholinguistic database (Coltheart, 1981): *child, earth, floor, glass, heart, board, house, mouth, blood, staff, plant,* and *teeth*. All were concrete (>400), high frequency (Kučera-Francis written frequency, also known as K-F-FREQ >90; Kučera & Francis, 1967), monosyllabic words five letters in length. As in Experiment 1, words were presented in an 8:1:1 (HI:LO:LO) ratio (see Fig. 1). This yielded a 2 (HI vs. LO contingency) x 2 (3 vs. 12 words) design, with contingency manipulated within-subject and number of words manipulated between-subjects.

4.1.4. Procedure

The procedure followed that of Experiment 1, except that words were now presented in uppercase. For the new 12-word condition, four words shared a high-contingency connection to each color (four times as many as standard; see Fig. 1).

4.1.5. Outlier removal

The data cleaning procedure was as described in Experiment 1. Appendix Table 1 lists the results of all data cleaning steps and shows the small amount of data trimmed. In the end, data from 163 of the 166 participants were used for our statistical analyses. A power sensitivity analysis using the *jpower* module for *jamovi* indicated that we were powered to detect between-subjects effect sizes in independent samples *t*-tests as small as $d = 0.44$ ($f = .22$, $\eta_p^2 = .05$) and within-subject effect sizes in paired-samples *t*-tests as small as $d = 0.31$ ($f = .16$, $\eta_p^2 = .02$), both with 80% power ($\alpha = .05$, two-tailed).

4.2. Results

The top half of Table 2 displays mean response time for correct responses on HI and LO trials in each group. A 2 × 2 mixed ANOVA was conducted, with Contingency (HI vs. LO) as the within-subject factor, Group (3-word vs. 12-word) as the between-subjects factor, and response time (in ms) as the dependent measure. As in Experiment 1, there was a significant main effect of Contingency, $F(1, 161) = 185.73$, $p < .001$, $\eta_p^2 = .54$, $BF_{10} > 100$, but not of Group, $F(1, 161) = 0.16$, $p = .693$, $\eta_p^2 < .01$, $BF_{01} = 2.31$, although the Bayesian evidence only reached anecdotal levels in the latter case. This time, however, the

Table 2
Experiment 2: response time and error rate.

Group	HI	LO	CL Effect
Response times			
3-word	540 (74)	579 (87)	39 (31)
12-word	547 (79)	561 (82)	14 (18)
Error rates			
3-word	.029 (.023)	.060 (.041)	.031 (.033)
12-word	.042 (.030)	.053 (.039)	.011 (.025)

Note. This table reports mean response time (in ms) and mean error rate (with standard deviations for each in parentheses) for high-contingency (HI) and low-contingency (LO) trials, and the mean contingency-learning (CL) effect (low contingency – high contingency) in the two experimental groups.

[4] We were, however, underpowered to reliably detect such effects if they were smaller than $d = 0.57$; see our power sensitivity analysis above.

Contingency x Group interaction was statistically significant, $F(1, 161) = 41.62$, $p < .001$, $\eta_p^2 = .21$, $BF_{10} > 100$: The contingency learning effect (LO – HI) was significantly larger for the 3-word group (39 ms). The effect was, however, statistically significant both in the 3-word group, $t(161) = 14.33$, $p < .001$, $d = 1.30$, $BF_{10} > 100$, and in the 12-word group, $t(161) = 5.03$, $p < .001$, $d = 0.79$, $BF_{10} > 100$. Planned comparisons showed that responses to HIs, $t(161) = 0.65$, $p = .517$, $d = 0.10$, $BF_{01} = 4.23$, and to LOs, $t(161) = 1.33$, $p = .186$, $d = 0.21$, $BF_{01} = 4.86$, were non-significantly different between groups (with moderate Bayesian evidence for the null model in each case).

4.3. Discussion

Experiment 2 conceptually replicated and extended the pattern observed in Experiment 1. As the number of words increased, the magnitude of the contingency learning effect diminished, contrary to the prediction of pure proportion/pure frequency accounts. We once again observed that the movement in group performance appeared to be mainly in the LOs becoming *faster* as more words were added (a statistically non-significant difference also apparent in Experiment 1).

Recall that the pure proportion/pure frequency accounts both predict that the HI trials should get slower as more words are associated with each color because there is a decrease in the frequency of instance episodes for *both* trial types while contingency ratios are maintained. Here, however, there was a 7 ms increase in response time to the HIs, as expected, but response time to the LOs was *reduced* by 18 ms. Although these differences were not statistically significant (likely due to inadequate statistical power to detect small between-subjects effects), the fact that the LOs actually sped up is not predicted by the pure proportion/pure frequency accounts. Moreover, the predicted overall slowing in responding with more words was again not evident.

5. Experiment 3

We conducted Experiment 3 including all three groups (3-word, 6-word, and 12-word) to replicate the previous findings in a single experiment.

5.1. Method

5.1.1. Participants

We used the same group-level sample sizes as in Experiment 1. A power sensitivity analysis is presented below in the Outlier Removal section. During the final academic term of 2018 and the first academic term of 2019, a total of 150 University of Waterloo undergraduate students took part in a single session in exchange for course credit. They were randomly assigned to condition, with 50 participants in each of the three conditions. All had self-reported normal or corrected-to-normal vision. After data trimming, the final sample of 148 participants used in our statistical analyses was 76% female, with age ranging from 17 to 31 ($M = 19.51$, $SD = 2.09$). Due to missing or incomplete data for some participants, these two demographic figures are based on 137 and 146 participants, respectively.

5.1.2. Apparatus

The apparatus was identical to that of Experiments 1 and 2.

5.1.3. Materials and design

The overall materials and design closely matched those of Experiment 2.[5] The experiment was a 2 (HI vs. LO contingency) x 3 (3 vs. 6 vs. 12 words) design, with contingency within-subject and number of words between-subjects.

5.1.4. Procedure

The procedure matched that of Experiment 2 except that we included all three levels of the shared associations: 3-word, 6-word, and 12-word.

5.1.5. Outlier removal

The data cleaning procedure again followed that in Experiment 1. Appendix Table 1 lists the results of all data cleaning steps for each experiment and the small amount of data trimmed. Here, data from 148 of the 150 participants were retained for use in our statistical analyses. A power sensitivity analysis using the *jpower* module for *jamovi* indicated that we were powered to detect between-subjects effect sizes in independent samples *t*-tests as small as $d = 0.57$ ($f = .28$, $\eta_p^2 = .08$) and within-subject effect sizes in paired-samples *t*-tests as small as $d = 0.40$ ($f = .20$, $\eta_p^2 = .04$), both with 80% power ($\alpha = .05$, two-tailed).

5.2. Results

The top half of Table 3 displays mean correct response time for HI and LO trials for each of the 3-, 6-, and 12-word groups across the first 150 trials (analyses of the full data set, presented in the Appendix, supported the same conclusions as presented here). As before, only response times for correct trials were analyzed. A 2 × 3 mixed ANOVA was conducted, with Contingency (HI vs. LO) within-subject, Group (3-word, 6-word, 12-word) between-subjects, and response time (in ms) from the first 150 trials in each group as the dependent measure.

As previously, the ANOVA revealed a significant main effect of Contingency, $F(1, 145) = 83.09$, $p < .001$, $\eta_p^2 = .36$, $BF_{10} > 100$, and a non-significant main effect of Group, $F(2, 145) = 2.23$, $p = .105$, $\eta_p^2 = .03$, $BF_{01} = 0.95$, although the Bayes factor for the latter result provided almost equal evidence for the null and alternative hypotheses. Consistent with our hypothesis, the Contingency x Group interaction was again significant, $F(2, 145) = 7.70$, $p < .001$, $\eta_p^2 = .10$, $BF_{10} = 34.1$, accompanied by strong Bayesian evidence as well. Welch-corrected

[5] There was one exception: To equate the number of presentations of individual stimulus combinations across all groups, we chose to vary the overall number of trials between groups. Consequently, there were 150, 300, and 600 trials presented in the 3-word, 6-word, and 12-word conditions, respectively. This resulted in an equal number of presentations for each HI or LO color-word combination across groups. Note that the ratio of HI:LO:LO trials was still 8:1:1 in each case. We subsequently decided that, because of the differing number of trials across groups, it was important to equate participant fatigue. Therefore, in the main text, we restrict our analyses to the first 150 trials of each group in Experiments 3 and 4. In the end, analyses based on all trials (see the Appendix) led to conclusions highly similar to those presented here, save for three differences. First, in Experiment 3, the LO vs. LO comparison between the 3-word and 12-word groups was significant when based on the first 150 trials but non-significant when based on the entire dataset. Second, the HI vs. HI comparison across groups in Experiment 4 was non-significant when based on the first 150 trials but significant when based on the entire dataset. Finally, the opposite was true for the LO vs. LO comparison in Experiment 4.

Table 3
Experiment 3: response time and error rate.

Group	HI	LO	CL effect
Response times			
3-word	561 (73)	607 (92)	46 (51)
6-word	541 (73)	564 (85)	22 (30)
12-word	549 (74)	567 (82)	18 (31)
Error rates			
3-word	.018 (.017)	.043 (.045)	.019 (.036)
6-word	.023 (.020)	.042 (.041)	.019 (.041)
12-word	.029 (.023)	.037 (.035)	.007 (.029)

Note. This table reports mean response times (in ms) and error rates for the first 150 trials (with standard deviations for each in parentheses) in high-contingency (HI) and low-contingency (LO) conditions, as well as the mean contingency-learning (CL) effect (low contingency – high contingency) in the three experimental groups.

independent-samples *t*-tests[6] revealed that the contingency learning effect was significantly larger for the 3-word group (46 ms) than for the 6-word group (22 ms), $p = .006$, and the 12-word group (18 ms), $p = .001$. The latter two groups, however, did not differ significantly, $p = .470$.

We again broke down the Contingency x Group interaction with planned comparisons, confirming significant contingency learning effects within each of the three groups: 3-word group, $t(145) = 8.41$, $p < .001$, $d = 0.91$, $BF_{10} > 100$; 6-word group, $t(145) = 4.07$, $p < .001$, $d = 0.74$, $BF_{10} > 100$; and 12-word group, $t(145) = 3.29$, $p = .001$, $d = 0.58$, $BF_{10} > 100$. Further planned comparisons exhibited significantly longer latencies for LOs in the 3-word group relative to the 6-word group, $t(145) = 2.49$, $p = .014$, $d = 0.49$, $BF_{10} = 2.80$, and the 12-word group, $t(145) = 2.31$, $p = .022$, $d = 0.46$, $BF_{10} = 2.11$, while Bayes factors provided only anecdotal evidence for the alternative in each case. Response times on LO trials in the latter two groups did not differ, $t(145) = 0.20$, $p = .845$, $d = 0.04$, $BF_{01} = 4.64$. For the HIs, no group difference was statistically significant ($ps \geq .190$).

We hypothesized that the size of the contingency learning effect would decline as more words were added in each subsequent condition: from 3 to 6 to 12 words. Because we hypothesized a monotonic decline in effect size, we conducted a Mann-Kendall trend test (also known as an M-K test; Kendall, 1938; Mann, 1945). We used the *mk.test()* function from the *trend* package for *R* to test the two-sided alternative that the RT effect size (LO - HI RTs) followed a monotonic trend moving from 3 to 6 to 12 words. Because the M-K test is non-parametric, it makes no assumptions based on normality of data, which suits RT data well.[7] This test revealed that the predicted monotonic decline in RT-based effect size was statistically significant, $S = -1597$, $\tau = -0.15$, $z = -2.65$, $p = .008$ (see Fig. 2).

5.3. Discussion

Experiment 3 replicated the patterns observed in Experiments 1 and 2. As the number of words increased, the magnitude of the contingency learning effect decreased in an orderly fashion. The interaction and the monotonic decline in performance in this experiment were both clear and significant. Consequently, we interpret the findings of this experiment as a successful conceptual replication of the previous two experiments. We saw a smooth decline in the contingency learning effect as the number of irrelevant words increased. There was again no overall slowing in responding as more words were added.[8]

That LOs became significantly *faster* as more words were added in the 6-word and 12-word groups, and that HIs were numerically slowest in the 3-word group, stands in opposition to the pure proportion/pure frequency accounts, as well as to the PEP 2.0 model, all of which predict general slowing of both HI and LO trials as more words are added.

Critically, according to the pure proportion/pure frequency accounts, responses on LO trials should never be expected to get *faster* as more words are added. Instead, these accounts both predict that LO responses should become slower—the opposite of what we have observed. Alternatively, one might expect LO responses to stay the same speed if they were already sufficiently rare that they did not receive facilitation they in the 3-word condition. But they should never get faster according to these 'facilitation-only' accounts. Alternative accounts that would predict speeded LO trials would include those that posit a role for response interference. Perhaps as HI trials become decreasingly prevalent in the trial set they exert less interfering influence on LO trial responses: Bias toward an incorrect, high-contingency color response for a given word is reduced when the learned contingency is less prevalent in the trial history.

With evidence in hand for such unpredicted differences in LO vs. LO comparisons between groups, we moved forward with a final experiment to replicate the most extreme disparity (3-word relative to 12-word) with a much greater sample size to achieve adequate power to detect potentially small between-group differences if they exist. Our goal in Experiment 4 was to confirm that as more words are added, the unpredicted decline in the contingency learning effect is in part—if not primarily—due to responses on LO trials speeding up rather than responses on HI trials simply slowing down. This pattern runs contrary to the predictions of the pure proportion/pure frequency accounts and instead is better aligned with the misprediction cost and bidirectional cost accounts that both suggest a role for response interference.

6. Experiment 4

In Experiments 1–3, the size of the contingency learning effect decreased as the number of words on the response-irrelevant dimension increased. However, it was not clear whether this reduction was due to slowing on HI trials, speeding on LO trials, or both. As noted earlier, the pure proportion/pure frequency accounts predict slowing of responding for both HI and LO trials as the number of words increases because the learned contingencies become diluted among so many unique color-word instances. This slowing should be substantial for HI trials, which are prevalent in the typical 3-word paradigm, but should also be apparent (albeit perhaps smaller) for LO trials which are rare in the 3-word paradigm and even rarer in the 12-word condition.

Of course, determining where the movement is in a between-subjects manipulation requires substantially greater statistical power than our experiments have had thus far, despite their already relatively large sample sizes. Therefore, the goal of Experiment 4 was to replicate Experiment 2 with a much greater sample size to provide sufficient statistical power to detect potentially small between-subjects movements separately for HI and LO trials. We predicted that if indeed response interference from HI trials typically is exerted on LO trials in the 3-word condition—as the misprediction cost and bidirectional cost accounts predict—then the dilution of HI trials in the 12-word condition may release LO trials from response interference, causing responses on LO trials to speed up. Put simply, if response interference is at play in contingency learning, we should see LO trials get faster in the 12-word condition; if no response interference is occurring, responses on LO trials

[6] All *t*-tests used throughout this manuscript were two-tailed with alpha set at .05.

[7] We used an M-K test here because we had no hypotheses concerning the shape of the trend. However, we conducted a linear regression as well. The result of this regression agreed with that of the M-K test, indicating that the size of the contingency learning effect was significantly negatively associated with Group, $R^2_{adj} = .08$, $F(2, 145) = 7.67$, $p < .001$.

[8] It is worth noting, however, that this experiment was underpowered to detect a between-subjects effect smaller than $d = 0.57$ (see our power sensitivity analysis above), so it is possible that there is a small yet significant group difference that we were unable to measure in this particular experiment.

should get slower or stay the same in the 12-word condition.

6.1. Method

6.1.1. Participants

An a priori power analysis was conducted using G*Power software (v. 3.1.9.7; Faul et al., 2007), aiming for the smallest effect size of interest that we could feasibly achieve power for. We chose to target Cohen's $d = 0.20$ for the critical LO vs. LO between-subjects comparison ($\alpha = .05$, two-tailed independent samples *t*-test) as measured by response times. This indicated required sample sizes of 788 or 1054 participants in total to achieve 80% or 90% statistical power, respectively. Accordingly, we aimed to collect a minimum of 788 participants in total, with our ideal target sample size set higher at 1054, although our stopping rule was based on the end of the final term in which we collected data. With the same parameters as listed above, it was also possible to detect a small between-subjects main effect ($f = .15$) of Group, which PEP 2.0 would predict based on the notion of overall slowing for HI and LO trials.

From the final academic term of 2020 until the final term of 2021, 1364 University of Waterloo undergraduate students each took part in a single session in exchange for course credit. All had self-reported normal or corrected-to-normal vision. The final sample of 1081 participants used in our statistical analyses after data trimming was 74% female with age ranging from 17 to 48 ($M = 20.47$, $SD = 3.60$). Due to missing or incomplete data for some participants, sex is based on 961 participants and age is based on 933 participants.

6.1.2. Apparatus

Unique to this experiment, participants completed the study on their own personal computers. We created a custom program that ran in a web browser and collected reliable response time data while being very similar in appearance to the in-lab versions of the program used in the previous experiments. The main experiment program was written primarily in Python, which was then converted to JavaScript and run locally in the participant's web browser (akin to an 'applet' style of program that services like PsychoPy offer). We hosted the program temporarily on participants' computers to minimize noise in response time collection. Data files were then uploaded to a university server after the participant had completed the study.

6.1.3. Materials and design

Again, participants were randomly assigned to one of two conditions, with 743 participants in the 3-word group and 621 in the 12-word group. The overall design closely followed that of Experiment 2 except that, as in Experiment 3, we chose to equate the number of presentations of individual stimulus combinations by varying the overall number of trials across conditions (see Footnote 3).

6.1.4. Procedure

The procedure was identical to that of Experiment 2, except that the experiment was completed online via a web browser on participants' own computers in their own space. This change also meant that response keys were no longer color-coded with stickers.

6.1.5. Outlier removal

The data cleaning procedure again followed that described in Experiment 1. Appendix Table 1 lists the results of all data cleaning steps for each experiment and the amount of data trimmed. We retained data from 1081 of the 1364 participants for our statistical analyses. Based on our a priori power analysis, this sample size ensures 90% power to detect a between-subjects effect in independent samples *t*-tests as small as $d = 0.20$. A power sensitivity analysis using the *jpower* module for *jamovi* confirmed that we were powered to detect between-subjects effect sizes in independent samples *t*-tests as small as $d = 0.20$ ($f = .10$, $\eta_p^2 = .01$) and within-subject effect sizes in paired-samples *t*-tests as small as $d = 0.13$ ($f = .07$, $\eta_p^2 = .004$), both with 90 % power ($\alpha = .05$, two-tailed).

As is shown in Appendix Table 1, most of the participant data removed here occurred during the second step of our standard data cleaning procedure—the step when data from participants with <80% accuracy across all trials were removed. The loss of data at this step is likely due to participants being more distracted than would be the case in the controlled environment of a laboratory. It is also possible that the absence of colored stickers on the keys, in contrast to previous experiments, was partly to blame for the higher error rate seen here.

6.2. Results

As in Experiment 3, we once again restricted all analyses to the first 150 trials in each group to mitigate any variance attributed to participant fatigue (see Footnote 3). The top half of Table 4 displays mean correct response times for HI and LO trials for the 3-word and the 12-word groups. Once again, only correct trials were analyzed. A 2 × 2 mixed ANOVA was conducted, with Contingency (HI vs. LO) within-subject, Group (3-word vs. 12-word) between-subjects, and response time (in ms) as the dependent measure. There was a significant main effect of Contingency, $F(1, 1079) = 373.04$, $p < .001$, $\eta_p^2 = .26$, $BF_{10} > 100$, indicative of a robust overall contingency learning effect. The main effect of Group was again non-significant, $F(1, 1079) = 1.66$, $p = .198$, $\eta_p^2 < .01$, $BF_{01} = 3.35$, with moderate Bayesian evidence for the null model. As in our prior experiments, the Contingency x Group interaction was significant, $F(1, 1079) = 110.14$, $p < .001$, $\eta_p^2 = .09$, $BF_{10} > 100$, demonstrating that the contingency learning effect was larger in the 3-word group (43 ms) than in the 12-word group (13 ms). Even with the change to an online setting, these means are very similar to those in Experiments 2 and 3. Planned comparisons confirmed significant contingency learning effects both in the 3-word group, $t(1079) = 22.37$, $p < .001$, $d = 0.81$, $BF_{10} > 100$, and in the 12-word group, $t(1079) = 5.91$, $p < .001$, $d = 0.33$, $BF_{10} > 100$.

The goal of this final experiment was to break down the significant Contingency x Group interaction to determine whether the reduction in contingency learning effect size is driven by a between-groups difference in the HIs, in the LOs, or in both. A planned comparison demonstrated that responses to HIs were non-significantly but numerically slower as more associations were added (i.e., in the 12-word group relative to the 3-word group; see Table 4), $t(1079) = 1.84$, $p = .067$, $d = 0.11$, $BF_{01} = 2.77$, although the Bayes factor only provided anecdotal evidence for the null model. Critically, and in contrast to the prediction from the pure proportion/pure frequency accounts, however, responses on LO trials were significantly *faster* in the 12-word group relative to the 3-word group, $t(1079) = 3.73$, $p < .001$, $d = 0.23$, $BF_{10} = 62.6$ (see Fig. 3), accompanied by very strong Bayesian evidence for this finding.

6.3. Discussion

Experiment 4 demonstrated that the attenuation of contingency learning with an increased number of response-irrelevant stimuli is

Table 4
Experiment 4: response time and error rate.

Group	HI	LO	CL effect
Response times			
3-word	574 (74)	617 (101)	43 (54)
12-word	583 (81)	596 (90)	13 (38)
Error rates			
3-word	.039 (.029)	.068 (.059)	.028 (.054)
12-word	.046 (.045)	.052 (.051)	.005 (.045)

Note. This table reports mean response times (in ms) and error rates for the first 150 trials (with standard deviations for each in parentheses) in high-contingency (HI) and low-contingency (LO) conditions, as well as the mean contingency-learning (CL) effect (low contingency – high contingency) in the two experimental groups.

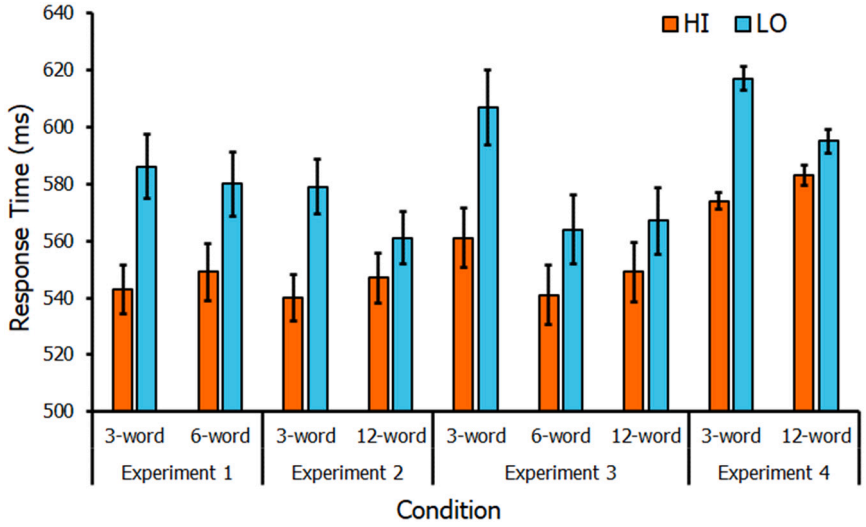


Fig. 3. Mean response times in all four experiments, split by HI and LO trials.
Note. Error bars = ± 1 *SE*.

highly reliable. These results were, however, at odds with predictions made by the pure proportion/pure frequency accounts, and by extension, by PEP 2.0 as well: the contingency learning effect should have stayed the same size as the number of stimuli on the response-irrelevant dimension increased, given that the contingency proportions and relative frequencies of HI and LO trails were constant. However, as we had noticed across Experiments 1–3, the results of Experiment 4 demonstrate that whereas HI responses became 9 ms *slower* as the number of response-irrelevant stimuli quadrupled between groups, LO responses became significantly *faster* by 22 ms. As a result, we again observed no overall slowing across groups, contrary to the prediction of facilitation-only explanations. Despite the first three experiments presented here being underpowered to detect small differences between groups, this final experiment, which could reliably measure main effects as small as $f = .15$, still did not find a significant main effect of Group (i.e.., overall slowing of HIs and LOs). Critically, that we observed significant speeding of LO responses as more words were added is consistent with the possibility of response interference influencing color-word contingency learning.

7. General discussion

Across four experiments, as an increasing number of words 'shared' contingencies with the same color, the learning of the color-word associations—the contingency learning effect—was progressively attenuated (see Fig. 2). We consistently replicated the prototypical contingency learning effect using a standard 3-word paradigm, where each word was most often presented (80% of the time) in a single color (one word per HI contingency). Further, we extended this to 6-word (2 words per HI contingency) and 12-word (4 words per HI contingency) versions. In so doing, we observed a systematic decline in the magnitude of the contingency learning effect as the number of connected words increased. This study represents the largest investigation of contingency learning to date with a total of 1490 participants retained for statistical analyses. Notably, Experiment 4, the basis of our critical conclusions, stands as the largest single contingency learning experiment yet conducted, featuring over 1000 participants.

Although other studies have investigated the effects of altering the contingency ratio (e.g., Forrin & MacLeod, 2018), we instead opted to maintain a consistent 8:1:1 ratio for HI:LO:LO color-word combinations. Without changing this contingency proportion, we systematically varied the number of words connected to a given color (e.g., in the 6-word group: both *mouth* in red; 8:1:1, and *under* in red; 8:1:1). Critically, this permitted the addition of more HI contingency combinations without reducing the overall proportion of LO trials (unlike what happens when the contingency ratio is altered). We now consider this consistent pattern of findings with reference to the five main accounts discussed at the outset of this article.

7.1. Facilitation-only versus interference-based accounts

At the beginning of this article, we summarized four main accounts that Schmidt and De Houwer (2016) had put forth as potential explanations for costs and benefits in the contingency learning effect. By way of reminder, they are: (1) prediction benefit, (2) misprediction cost, (3) bidirectional cost, and (4) pure proportion. Later, we summarized a fifth account that shares many qualities with the pure proportion account—the pure frequency account. It is important to recognize that whereas the prediction benefit, pure proportion, and pure frequency accounts suggest that contingency learning is entirely facilitative, the misprediction cost and bidirectional cost accounts instead posit a role for response interference (i.e., expectations that are set out by experience with HIs make all other response options slower).

Critically, proponents of facilitation-based accounts should only ever predict faster performance for LOs under two highly similar scenarios: (1) LOs share a higher *proportion* of all trials such that there are proportionately more matching instances in memory to point at the current trial, or (2) LOs become more *frequent* in terms of their raw trial count such that, again, more matching instances in memory point at the current trial. Neither of these scenarios was true in our study. Instead, the number of unique items presented to participants was manipulated without changing the overall contingency ratios and without increasing the frequency of particular instances. This alteration actually *decreased* the frequency of any given HI or LO item, since there were more unique combinations to encode. These changes should have caused responses on LO trials either to stay the same (according to the pure proportion account) or to become slower (according to the pure frequency account).

Our findings implicate an additional factor at play in contingency learning, beyond any effects of biased facilitation. In the face of decreased frequency and stable contingency proportion, the only way that a response to a LO trial can get *faster* is if it is no longer being 'held back' by interference from HI trials. Of the five accounts presented by Schmidt and De Houwer (2016), only the misprediction and bidirectional cost accounts posit roles for response interference. Therefore, we put those two accounts forward as the best fit for the present findings. Next, we consider what the inclusion of response interference would

mean for PEP.

7.2. Implications for the PEP 2.0 model of contingency learning

Recall that PEP 2.0 is a computational model of stimulus-response bindings that implements the principles of instance theories to explain performance changes in a variety of cognitive phenomena (contingency learning, Stroop, stimulus-response bindings effects, mixing costs, etc.). It argues two basic premises: (1) each individual event is encoded and stored as a separate episode in memory, and (2) during subsequent events, stored episodes are retrieved simultaneously to aid performance by biasing response selection. Since in contingency learning there are many more HI episodes in memory, retrieval is biased toward a large facilitative response on HI trials and a small facilitative response on LO trials.

The PEP 2.0 model most closely resembles a 'pure proportion' account, and consequently, would predict a main effect of Group but no interaction in our experiments: Overall performance should get slower as more words are added to the set (i.e., due to diminished facilitation for all trials as a result of their increased scarcity) but the size of the effect should remain the same. That we observed precisely the opposite pattern of results across four experiments (an absent main effect of Group and a significant interaction) certainly contradicts the model's predictions. Of course, this lack of a main effect of word set size may be at least partially explained by the LOs actually getting *faster* while the HIs got predictably slower, effectively cancelling each other out (see, for example, Experiment 4 represented in Fig. 3). Together, these two changes—HIs slowing down (or holding stable) and LOs speeding up—resulted in an overall reduction in the contingency learning effect. That a substantial portion of this change stemmed from the LO trials becoming faster with increased number of associations runs counter to the PEP 2.0 model: The model would never predict *facilitation* of performance for HI or LO trials when episodes are 'watered down' with more color-word combinations.

Accounts incorporating response interference are consistent with a previous claim (Lin & MacLeod, 2018) that the contingency learning effect is made up of both benefits for HI trials and costs for LO trials. As additional evidence for this claim, consider that if LOs are typically slowed by interference from HIs, it stands to reason that the HI associations need to be learned first, before they can exert any response interference on LOs. Indeed, previous work (Lin & MacLeod, 2018) has already found support for this notion: The benefit for HI trials is quickly realized whereas the cost to LOs takes longer to appear.

In summary, we reason that the size of the contingency learning effect decreased in our experiments because of two related occurrences: HIs slowed down—due to decreased facilitation—and LOs sped up—due to decreased interference exerted by those HIs. Contingent on the proportion of stored episodes, one may observe varying degrees of response facilitation. PEP 2.0 handles this well. Arguably, these stored episodes can, however, also prime individuals for a response, necessitating additional work to overcome this interference so as to make the correct response. As a result, response interference should be considered as a component mechanism in the next iteration of the PEP model. In fact, facilitation and interference need not be mutually exclusive ideas: Responses to HIs may be facilitated due to the increased number of instances in memory, as PEP 2.0 would predict, whereas LOs may not only lack facilitation due to there being few instances in memory (in accordance with the PEP model), but they may also suffer additional interference from HIs (in accordance with response interference).

7.3. Limitations

While our interpretation of the data thus far has centered on theory-relevant mechanisms of facilitation and interference, it remains possible that an unintended group dynamic is in play. For instance, the 12-word version of the task, with its considerably higher variety of unique trials, is perhaps more engaging to participants.[9] This could have reduced response times for both HI and LO trial-types. Simultaneously, a facilitation-only hypothesis would predict that the benefit for HI trials would be reduced due to their increased scarcity, leading to slower response times that offset the benefit derived from increased focus. In the end, such a pattern of results would be similar to what we found here: LO trials would become faster while HI trials would remain relatively stable. While plausible, this 'increased engagement' explanation is unlikely because the task is held constant between groups; participants must respond to the color of the word regardless of how many unique color-word combinations are presented.

Some have argued that in the color-word contingency learning paradigm, the response-irrelevant word is mapped to a response key rather than to the response-relevant dimension (i.e., the font color), and thus initial processing of a word may facilitate the correct key response even before the color is identified (Schmidt et al., 2007). This change in the locus of association binding however, is inconsequential to our findings. In either case, the critical factor is that performance is facilitated for HI trials and that built up 'expectations', whether about the color or the associated keypress, then slow responses on LO trials—a form of interference.

Perhaps more importantly, it is worth acknowledging that our critical results were statistically significant but relatively small: the LO vs. LO comparison to yielded a small effect size of $d = 0.23$ in Experiment 4. In addition, while this LO vs. LO difference was numerically present in all four experiments, it only emerged as statistically significant when our sample size exceeded 1000, and in a design featuring an extreme disparity between 3- and 12-word conditions. This may be more a consequence of the high statistical power required for detecting small between-subjects effects, and therefore it may be possible to observe the effect of speeded LO responses in clever within-subject designs with fewer participants. Therefore, while response interference may not play a large role in simple association learning, it is nevertheless a factor that must be considered because it can account for detectable variance in contingency learning.

8. Conclusion

This study is the first to demonstrate the effect of associating multiple items on a response-irrelevant dimension (words) to a single response-relevant dimension (color) in contingency learning. Across four experiments, adding more words as highly contingent with a given color systematically reduced the contingency learning effect, contrary to predictions by the PEP 2.0 model (Schmidt et al., 2016) and related facilitation-only models. Increasing the number of specific color-word instances 'waters down' encoded episodes, diminishing the likelihood that matching instances are in recent memory, and ultimately decreasing the influence that prior learning has on the current trial. This leads to slower responding for high-contingency associations because the benefit of facilitation is diminished. We have suggested that this also leads to *faster* responding to low-contingency associations that now suffer reduced response interference from diminished expectations stemming from high-contingency trial episodes. Incorporation in PEP 2.0 of an additional mechanism—interference imposed on LOs caused by HIs—provides one way to capture the currently unpredicted speeding of low-contingency episodes. At the same time, this would account for overall responding not slowing down as the number of response-irrelevant instances increases. We argue that potential contributions from response interference should be considered not only in the contingency learning paradigm, but also in related studies examining learning of probabilities and simple associations.

[9] We thank Reviewer 1 for this suggestion.

Author note

This research was supported by a Natural Sciences and Engineering Research Council (NSERC) of Canada Postgraduate Scholarship to BRTR and by NSERC Discovery Grant A7459 to CMM. An early version of this work was presented at the 59th meeting of the Psychonomic Society in 2018. We thank Torin Young and Kristen Sullivan for their assistance in collecting and organizing participant data. We also thank Eric Lavigne and Chris Xu for developing the program used in Experiment 4. Our data, analysis code, experiment programs, and other materials are listed on the Open Science Framework (OSF; https://osf.io/4x9v8/).

CRediT authorship contribution statement

Brady R.T. Roberts: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing. **Noah D. Forrin:** Conceptualization, Investigation, Methodology, Writing – review & editing. **David McLean:** Data curation, Investigation, Project administration, Resources, Software, Validation, Writing – review & editing. **Colin M. MacLeod:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – review & editing.

Declaration of competing interest

The authors have no conflict of interest to declare.

Data availability

All program files, data, and statistical analysis code are available on the Open Science Framework (OSF; https://osf.io/4x9v8/).

Appendix A. Supplementary analyses

Supplementary analyses to this article can be found online at https://doi.org/10.1016/j.actpsy.2024.104187.

References

Besner, D., McLean, D., & Young, T. (2021a). Do eyes and arrows elicit automatic orienting? Three mutually exclusive hypotheses and a test. *Quarterly Journal of Experimental Psychology, 74*(7), 1164–1169. https://doi.org/10.1177/1747021821998572

Besner, D., McLean, D., & Young, T. (2021b). On the determination of eye gaze and arrow direction: Automaticity reconsidered. *Canadian Journal of Experimental Psychology, 75*(3), 261–278. https://doi.org/10.1037/CEP0000261

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods, 51*(2), 467–479. https://doi.org/10.3758/s13428-018-1077-9

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A, 33*(4), 497–505. https://doi.org/10.1080/14640748108400805

E-Prime (v. 3.0). (2016). *Psychology Software Tools Inc. [computer software]*.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Forrin, N. D., & MacLeod, C. M. (2018). Contingency proportion systematically influences contingency learning. *Attention, Perception, and Psychophysics, 80*(1), 155–165. https://doi.org/10.3758/S13414-017-1424-4

Geukes, S., Vorberg, D., Zwitserlood, I., & P.. (2019). Disentangling semantic and response learning effects in color-word contingency learning. *PLoS One, 14*(5), Article e0212714. https://doi.org/10.1371/JOURNAL.PONE.0212714

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist, 73*(1), 2. https://doi.org/10.1037/AMP0000263

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*(1–2), 81–93. https://doi.org/10.2307/2332226

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: An integrative review. *Psychological Bulletin, 142*(10), 1111–1140. https://doi.org/10.1037/BUL0000064

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Lenth, R. V., Buerkner, P., Herve, M., Jung, M., Love, J., Miguez, F., … Singmann, H. (2022). *emmeans: Estimated marginal means, aka least-squares means (version 1.8.1-1). [computer software]*.

Lewicki, P. (1985). Nonconscious biasing effects of single instances on subsequent judgments. *Journal of Personality and Social Psychology, 48*(3), 563–574. https://doi.org/10.1037/0022-3514.48.3.563

Lin, O. Y.-H., & MacLeod, C. M. (2018). The acquisition of simple associations as observed in color-word contingency learning. *Journal of Experimental Psychology: Learning Memory and Cognition, 44*(1), 99–106. https://doi.org/10.1037/XLM0000436

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*(4), 492–527. https://doi.org/10.1037/0033-295X.95.4.492

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82*(4), 276–298. https://doi.org/10.1037/h0076778

MacLeod, C. M. (2019). Learning simple associations. *Canadian Psychology, 60*(1), 3–13. https://doi.org/10.1037/cap0000170

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica, 13*(3), 245–259. https://doi.org/10.2307/1907187

Morey, R. D., & Selker, R. (2020). *jpower: Power analysis for common research designs (version 0.1.2) [computer software]*.

Pohlert, T. (2020). *Trend: Non-parametric trend tests and change-point detection* (version 1.1.4). [computer software].

R Core Team. (2020). *R: A language and environment for statistical computing (version 4.1.1). R Foundation for Statistical Computing. http://www.r-project.org/ [Computer software]*.

Şahin, M., & Aybek, E. (2019). Jamovi: An easy to use statistical software for the social scientists. *International Journal of Assessment Tools in Education, 670–692*. https://doi.org/10.21449/ijate.661803

Schmidt, J. R. (2016). Proportion congruency and practice: A contingency learning account of asymmetric list shifting effects. *Journal of Experimental Psychology: Learning Memory and Cognition, 42*(9), 1496–1505. https://doi.org/10.1037/XLM0000254

Schmidt, J. R. (2018). Best not to bet on the horserace: A comment on Forrin and MacLeod (2017) and a relevant stimulus-response compatibility view of colour-word contingency learning asymmetries. *Memory & Cognition, 46*(2), 326–335. https://doi.org/10.3758/s13421-017-0755-7

Schmidt, J. R. (2021). Incidental learning of simple stimulus-response associations: A review of colour-word contingency learning research. *L'Année Psychologique, 121*(2), 77–127. https://doi.org/10.3917/ANPSY1.212.0077

Schmidt, J. R., Augustinova, M., & De Houwer, J. (2018). Category learning in the color-word contingency learning paradigm. *Psychonomic Bulletin & Review, 25*(2), 658–666. https://doi.org/10.3758/s13423-018-1430-0

Schmidt, J. R., Crump, M. J. C., Cheesman, J., & Besner, D. (2007). Contingency learning without awareness: Evidence for implicit control. *Consciousness and Cognition, 16*(2), 421–435. https://doi.org/10.1016/J.CONCOG.2006.06.010

Schmidt, J. R., & De Houwer, J. (2012a). Does temporal contiguity moderate contingency learning in a speeded performance task? *Quarterly Journal of Experimental Psychology, 65*(3), 408–425. https://doi.org/10.1080/17470218.2011.632486

Schmidt, J. R., & De Houwer, J. (2012b). Adding the goal to learn strengthens learning in an unintentional learning task. *Psychonomic Bulletin & Review, 19*(4), 723–728. https://doi.org/10.3758/S13423-012-0255-5

Schmidt, J. R., & De Houwer, J. (2012c). Learning, awareness, and instruction: Subjective contingency awareness does matter in the colour-word contingency learning paradigm. *Consciousness and Cognition, 21*(4), 1754–1768. https://doi.org/10.1016/J.CONCOG.2012.10.006

Schmidt, J. R., & De Houwer, J. (2016). Contingency learning tracks with stimulus-response proportion: No evidence of misprediction costs. *Experimental Psychology, 63*(2), 79–88. https://doi.org/10.1027/1618-3169/a000313

Schmidt, J. R., De Houwer, J., & Besner, D. (2010). Contingency learning and unlearning in the blink of an eye: A resource dependent process. *Consciousness and Cognition, 19*(1), 235–250. https://doi.org/10.1016/J.CONCOG.2009.12.016

Schmidt, J. R., De Houwer, J., & Rothermund, K. (2016). The Parallel Episodic Processing (PEP) model 2.0: A single computational model of stimulus-response binding, contingency learning, power curves, and mixing costs. *Cognitive Psychology, 91*, 82–108. https://doi.org/10.1016/j.cogpsych.2016.10.004

Seligman, M. E. (1970). On the generality of the laws of learning. *Psychological Review, 77*(5), 406–418. https://doi.org/10.1037/H0029790

Shanks, D. R. (2007). Associationism and cognition: Human contingency learning at 25. *Quarterly Journal of Experimental Psychology, 60*(3), 291–309. https://doi.org/10.1080/17470210601000581

Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M. B., Højsgaard, S., … Christensen, R. H. B. (2022). *afex: Analysis of factorial experiments (version 1.1-1). [computer software]*.