Trial frequency outweighs trial duration in associative learning: Generality and boundary conditions

Journal:	*Quarterly Journal of Experimental Psychology*
Manuscript ID	QJE-STD-24-321.R2
Manuscript Type:	Standard Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Witnauer, James; SUNY Brockport, Psychology Chew, Sarah; SUNY Binghamton Powell, Jennifer; SUNY Binghamton Murphy, Robin; University of Oxford, Experimental Psychology Miller, Ralph; State University of New York at Binghamton, State University of New York at Binghamton; Binghamton University (SUNY),
Keywords:	associative learning, contingency, trial spacing effect, trial frequency, trial duration

Trial frequency outweighs trial duration in associative learning:

Generality and boundary conditions

James E. Witnauer[1], Sarah Chew[2], Jennifer Powell[2], Robin A. Murphy[3], and Ralph R. Miller[2]

[1]State University of New York - Brockport

[2] State University of New York - Binghamton

[3]University of Oxford

Contact information:
James E. Witnauer
Department of Psychology
State University of New York – Brockport
Brockport, NY, 14420 USA
jwitnaue@brockport.edu

Keywords: associative learning, contingency, trial spacing effect, trial frequency, trial duration

Author notes: All experiments were approved by the SUNY-Binghamton Institutional Review Board. The authors thank Dennis Elengickal and Nathaniel Darko for assistance in programming and Adrianna Agnello, Rafi Arnob, Kimberly Casey, Edward Cook, Dave Jiang, Edeline Kalishevich, Lucas Petruzzo, and Denis Pogosyan for commenting on a draft of the manuscript. The research was supported in part by NIH grant MH033881. Raw data and statistical analyses are available at https://osf.io/vznr4/?view_only=a7182cf8c13942d49ce011f656446c1e

Abstract

Perceived contingency of a single cue and outcome is based on the relative exposure to four types of events: Cue-outcome pairings (A events), cue-alone presentations (B events), outcome-alone presentations (C events), and events in which neither the cue nor the outcome is presented (D events). Previous experiments found increases in the frequency of an event affected ratings of the perceived contingency between the cue and outcome, even compared to conditions with proportional decreases in the duration of trials (i.e, adjusted frequency conditions). The present experiments tested the generality and boundaries of this adjusted frequency effect by examining whether it generalizes to ratings of multiple cue-outcome dyads, to a cued-recall test, and to both sequential and simultaneous cue-outcome presentations. Experiment 1 revealed a strong effect of frequency but no effect of duration after training with a single cue-outcome dyad; however, a duration effect emerged when training consisted of five cue-outcome dyads. Experiment 2 showed an effect of duration as well as an adjusted frequency effect in contingency ratings after training with five dyads. Experiment 3 extended these observations to a cued-recall test after training with ten cue-outcome dyads. Experiment 4 used five dyads and found a within-experiment effect of duration on both contingency ratings and cued-recall scores. Whereas Experiments 1-4 varied the A events, Experiment 5 varied frequency and duration of the D events with ten cue-outcome dyads and revealed effects of duration as well as frequency on both cued recall and cue-outcome contingency ratings. In summary, these experiments detected an increase in the importance of event duration with increases in the number of dyads. Moreover, subject ratings of contingency closely tracked results in a cued-recall test, suggesting that a common mechanism underlies these two measures.

Trial frequency outweighs trial duration in associative learning:

Generality and boundary conditions

Associative learning is a change in the connection between mental representations of stimuli. In the excitatory case, a later presentation of one stimulus alone results in activation of the representation of a second, absent stimulus. Pavlovian conditioning with a conditioned stimulus and an unconditioned stimulus illustrates such an associative effect (e.g., Pavlov, 1927). In addition, human cognitive processes seem to depend on associative mechanisms at the time of initial encoding and retrieval, as demonstrated in human contingency learning (Shanks, 1985), category learning (Gluck & Bower, 1988), paired-associate learning (e.g., Bower, 1962), cued-recall (Aue et al., 2012), and associative recognition (Criss & Shiffrin, 2004). Given the generality of associative learning, it is important to identify the procedural variables that affect the strength of associative learning. In previous experiments (Murphy et al., 2022), we have found that a streamed trials procedure (Crump et al., 2007) is useful for studying these variables because the short total duration of each experimental condition facilitates the use of fully within-subjects designs with potentially dozens of different experimental conditions. The streamed trial procedure exposes participants to a rapidly presented 'stream' of associative events constituting an experimental condition, which produces strong contingency ratings by participants despite the use of extremely short-duration stimuli. For example, event durations in Crump et al.'s original experiments were only 100 ms; yet, participants perceived not only the individual events but also how the two stimuli were related. This invites the question: Does event duration matter in the streaming procedure?

Previous experiments from our laboratories have varied the frequency and duration of events and generally reported stronger effects of the frequencies of events (e.g., cue-outcome

pairings) than the durations of events on judgments of the contingency between a cue and an outcome, despite multiplicatively equivalent changes in event frequency and duration (Castiello et al., 2022; Murphy et al., 2022; Witnauer et al., 2023). Participants' judgments of the contingency between a cue and an outcome were affected by the frequencies of four different event types:

A events consisting of pairings of a cue with an outcome.

B events consisting of presentations of the cue alone.

C events consisting of presentations of the outcome alone.

D events consisting of cue-outcome co-absence.

The effect of frequency of a given event type was stronger than the effect of duration of that event type. The small effect of event duration on contingency judgments relative to the effect of event frequency results in a *free lunch* effect in which the impact of an event type on contingency ratings increases with increases in the frequency of the event, even when event duration is decreased proportionally, thereby equating total duration of exposure to the event across conditions (*free lunch* as in 'something for nothing,' costing nothing in time; Murphy et al., 2022). For example, a four-fold increase in the number of A events increases ratings of the contingency between the cue and outcome, even when the duration of each A event is decreased by a factor of four. In other words, manipulations of frequency affect ratings even when the manipulation includes adjustment of the duration of the event that was inversely proportional to the adjustment of frequency. This is a potentially important observation because it suggests a means of accelerating learning without increasing the duration of the training session and, in the extreme case, enhancing learning while actually shortening the duration of the training session

(e.g., increasing the frequency of a given event type by a factor of two while decreasing the duration of that event type by a factor of four).

One important limitation of previous experiments demonstrating the free lunch effect is that they have used only subjective ratings of relatedness of the two events (i.e., contingency judgments). Here, we sought to extend the measures of memory that might exhibit the free lunch effect. The present experiments were concerned with whether memory of an event is improved by using a large number of brief exposures to that event relative to a small number of long-duration exposures. Specifically, we tested whether the free lunch effect occurs when participants have a higher cognitive load than that provided by a single cue and outcome. Toward this end, we asked participants to learn about multiple cue-outcome dyads within a single stream of training events. Human memory is affected by list length (i.e., the number of pairs of items in a to-be-recalled list of associates). That is, memory for any single associate usually decreases with increases in the number of to-be-remembered pairs of items. Measurements of memory typically include accuracy (e.g., d'), proportion of test items correctly retrieved, or total number of correctly recalled items (Brady, Robinson, Williams, et al., 2023), with these different measures sometimes diverging. In simple list learning, increases in list length tend to decrease the percent of items people can recall from the list while simultaneously increasing the number of items recalled. For example, if training consists of either 10 or 20 items, correctly recalling 8 items out of 10 (80% correct) would constitute a greater proportion correct than recalling 14 items correctly out of 20 (70% correct).

This raises the question of why, if people can recall 14 items, they cannot recall more than 8 with shorter lists. In free recall situations, increases in list length produce similar opposing changes in proportion correct and number correct (Murdock, 1968). For example, Roberts (1972)

conducted a free recall experiment in which participants received lists containing 10, 20, 30, or 40 items. Increases in the number of items resulted in increases in the number of correctly recalled items and decreases in the proportion of items correctly recalled. The total-time hypothesis proposes that the level of recall (i.e., the proportion of correctly recalled items) after a study period will be a linear function of the total amount of time spent studying and the number of items in the list (Murdock, 1960). In addition, equivalent recall levels should be achieved when increases in list length are accompanied by a proportional reduction in study time per item. For example, equal recall levels should be achieved after exposure to a list of 20 words at 2 sec each and 40 words at 1 sec each because the two conditions use the same total amount of study time. Although most of this research measured free recall, similar principles appear to apply to cued-recall (e.g., Unsworth & Engle, 2007). It remains unknown whether the adjusted frequency effect would generalize to 1) a cued-recall procedure or 2) a procedure in which participants track multiple dyads.

The present experiments were concerned with the generality of the adjusted frequency (free lunch) effect. Specifically, after demonstrating the free lunch effect once again with contingency ratings in Experiments 1 and 2, we sought to demonstrate the free lunch effect using a cued-recall test of memory for cue-outcome dyads presented during a trial stream. We selected a cued-recall procedure in part because this procedure would assess generalization of the effect and has greater applied value than tasks that test contingency ratings alone. For example, cued-recall provides a laboratory model of measurements of vocabulary acquisition (e.g., Avila & Sadoski, 1996). Centrally, measurement of cued-recall would speak to the generality of the free lunch effect beyond contingency ratings. The present experiments also used multiple cue-outcome dyads during training in order to further evaluate the generality of the free lunch effect.

At test, participants were either asked to rate the contingency between two presented images or presented with the cue and asked to name the previously paired outcome by typing a response into a textbox.

In five experiments using the streamed trial procedure, participants received experimentally varied levels of exposure to various event types. Before testing whether the free lunch effect occurs with cued-recall, we attempted to replicate the free lunch effect for A and B events with contingency ratings when the participant was exposed to multiple cue-outcome dyads as opposed to a single cue-outcome dyad in each condition (Experiments 1 and 2). Experiment 3 tested whether the free lunch effect occurs for A events in cued-recall after training with multiple dyads. Experiment 4 directly compared cued-recall and contingency ratings after training with multiple dyads in which exposure to A events was manipulated across conditions. Experiment 5 tested for the free lunch effect for D events using both cued-recall and contingency tests after training with multiple dyads.

Exposure was manipulated by changing the frequency, duration, or adjusted frequency of an event in a way that allowed detection of the free lunch effect. Variations in exposure in a stream were followed either by a contingency rating test, a cued-recall test, or both. In addition, each target cue-outcome dyad was presented in a stream that consisted of either one, three, five, or ten cue-outcome dyads. All five of the present experiments are summarized in Table 1. The manipulated variables, measurements, stimuli, and levels of event exposure were selected in a way that allowed us to test whether the free lunch effect would generalize to new situations with greater applied value and demands on cognitive load than our previous experiments (Murphy et al., 2022). Based on the enhanced opportunity for interference between dyads, there is strong

reason to think that learning about multiple dyads in a condition would produce, if nothing else, weaker performance per dyad than a condition that involved only one dyad (e.g., Bower, 1962).

Experiment 1

Experiment 1 tested the effects on contingency ratings of frequency and duration of A and B cell events with streamed training consisting of one, three, or five cue-outcome dyads. Dyad number (One vs. Three vs. Five) and event type manipulated (A vs. B) were fully factorial. In addition, A (or B) event duration (150 ms vs. 2400 ms) and frequency (3 vs. 48 for each dyad) were manipulated by four-fold changes relative to a common baseline condition that included 12 presentations at 600 ms of each event type (per dyad for cells A, B, and C). Increasing the number of dyads across otherwise equivalent conditions was accompanied by an increase in the total number of A, B, and C events while holding constant the number of presentations of any specific dyad. All conditions included 12 D events. See Table 2 for a summary of the design of Experiment 1. Thus, Experiment 1 consisted of 27 conditions. Based on the results of previous experiments, we expected to observe a strong effect of frequency and a weak effect of duration in the One Dyad conditions, which constituted a conceptual replication of the central results of Murphy et al. (2022). The design also permitted us to test whether the relative insensitivity to duration would change when training consisted of three or five dyads.

Methods

Transparency and openness statement for all experiments

The procedures used for making data and code available to readers were the same for all experiments reported in this paper, except where otherwise stated within the Methods sections of specific experiments. The raw data and R code used to analyze data are available through the

Open Science Framework at https://osf.io/vznr4/?view_only=a7182cf8c13942d49ce011f656446c1e. The programs for these experiments are available at

Experiment 1: https://app.gorilla.sc/openmaterials/871358.

Experiment 2a: https://app.gorilla.sc/openmaterials/871364.

Experiment 2b: https://app.gorilla.sc/openmaterials/871365.

Experiment 3: https://app.gorilla.sc/openmaterials/871368.

Experiment 4: https://app.gorilla.sc/openmaterials/871369.

Experiment 5: https://app.gorilla.sc/openmaterials/871370.

Participants

Fifty-nine participants, consisting of 16 females and 43 males, were recruited from Amazon's Mechanical Turk crowdsourcing platform to be run in a 50-min session on a Gorilla Experiment Builder program. In principle, fifty-nine participants resulted in sensitivity to small effects ($f = 0.10$) with 80% statistical power in a design that consists of 15 experimental conditions (within subjects), assuming that the correlation between repeated measurements is $r = .50$. For consistency across experiments, we aimed to recruit 59 participants in each experiment reported in the present paper. The average age was $M = 38$ years ($SD = 7$). Recruitment was restricted to a population that included Mechanical Turk workers who had achieved approval of over 99% on at least 500 previous tasks and who were participating in the AU, CA, NZ, UK, and US using Chrome, Firefox, or Safari web browsers. Participants were required to complete the task within 3 hours of accepting the task. Participants were compensated with US $6 for their

time. Participants who reported being older than 50 years or younger than 18 years were precluded from starting the study. Individuals who are older than 50 (and younger than 18) tend to show greater sensitivity to photogenically-induced seizures. Hence, they were excluded for ethical reasons. All of a participant's scores were excluded if there was no variability in their responses across conditions (e.g., a participant entered a rating of 8 for all conditions). One participant was excluded for this reason. Based on prior research, a participant's response to a condition was excluded if they took more than 30 seconds to respond to the test question. Twenty-eight participants provided incomplete data, with at least one of the conditions having a reaction time of greater than 30 seconds. A 'missing completely at random' test in R (Yanagida, 2023) was used to test whether there was a pattern in the missing data. This test failed to reach significance, $\chi^2(632) = 593.73, p = .86$.

Stimuli

Eighty-one unique pairs of cue and outcome images served as dyads in the present experiment. Figure 1 presents a template for the placement of the cue and outcome images, which were always 130- x 130-pixel images surrounded by thick, black 240- x 190-pixel rectangular frames. Cue-outcome dyad membership was fixed such that each cue was paired with only one outcome, but dyads were randomly assigned to conditions across participants. A frame was also presented around the empty space that was created by the omission of an outcome (B cell events), cue (C events), or both (D events). Cues were always presented in the top half of a computer screen and consisted of either a Greek letter, Arabic letter, Anglo-Saxon character, Oracle bone script character, or occult symbol. Outcomes were always presented in the lower half of the screen and consisted of line drawings of common objects that could be easily named.

Procedure

The experiment was programmed in Gorilla Builder 1 (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020) and conducted using the Gorilla platform. Upon selecting the task, participants clicked a link to navigate from Mechanical Turk to the Gorilla website. After providing informed consent, participants provided demographic information, reporting their gender and year of birth. Next, participants were asked to report the kind of computer that they were using to complete the task. Tablet and smartphone users were excluded from the study.

After reading the instructions that provided an overview of the task and an explanation of how to use the contingency rating scale (see Appendix A), participants were asked to prepare to engage the task without interruption during each stream of trials. When participants finished reading the instructions, the first experimental condition was presented. The order of conditions was randomized anew for each participant. The sequence of events within a condition was determined by block randomization such that the contingency within each block was identical to the overall contingency of the condition. There were three blocks of events within each condition. Randomization of events within a block was constrained such that trials alternated in their left-right position on the screen relative to a central fixation cross. This ensured that when two events of the same type were presented back-to-back, participants would be able to detect that two events (instead of one long event) had occurred. All stimulus elements (i.e., cue and outcome) were presented simultaneously during an event. Hence, the terms ‘cue’ and ‘outcome’ are used here only for convenience and have no meaning with respect to temporal order during the training streams. Immediately after the last trial of each condition, one dyad was rated for contingency. The last event of the condition was immediately replaced by a rating screen. Text in the top-left quadrant of the rating screen asked participants to respond as quickly as possible.

Please indicate the degree of relatedness between these 2 pictures in the series you just saw. Make sure to answer as quickly as possible to stay in the experiment.

The top right quadrant of the screen contained an image with both members of one of the dyads that the participant received during the stream (cue = top, outcome = bottom). In the Three and Five Dyad conditions, this dyad was randomly selected. Below that image was a slider with anchors at -10, 0, and +10 for the relatedness of the cue and outcome. The initial position of the slider was always zero. The slider was accompanied by the text: *Use the rating scale below to enter the degree of relatedness.*

Event type. In the A event conditions, either the frequency or duration of A events was increased or decreased by a factor of four relative to the baseline frequency (12/dyad) or duration (600 ms), and the frequency and duration of B events were held constant at the baseline values (frequency = 12/dyad, duration = 600 ms). This was done for all dyads in a given condition. Thus, the frequency of A events for each of the one, three, or five dyads was manipulated in the Few A and Many A conditions, and the duration of A events for each of the one, three, or five dyads was manipulated in the Short A and Long A conditions. Similarly, across the B conditions, either the frequency or duration of B events was manipulated by a factor of four relative to the baseline frequency (12/dyad) or duration (600 ms), and the frequency and duration of A events were held constant at the baseline values (frequency = 12/dyad, duration = 600 ms).

Dyad number. Across conditions, the number of dyads (1, 3, or 5) was manipulated orthogonally to the other independent variables. In baseline conditions, the baseline number of A, B, and C events (12) was presented for each of the dyads, plus 12 D events. Thus, all One Dyad conditions involved 12 C and 12 D events. In all Three Dyads conditions, 36 C events (12

for each dyad) and 12 D events were presented in addition to the number of A and B events determined by condition assignment as described below. In all the Five Dyads conditions, 60 C events and 12 D events were presented. Thus, all conditions included 12 D events, whereas the number of C events increased along with the number of dyads.

Event duration. For those events specified by the Event Type factor, the duration of the event type was manipulated by a factor of four relative to the baseline. The Short conditions received 150-ms presentations of either A or B events. The Long conditions received 2400-ms presentations of either A or B events. The durations of all other events (B or A, as well as C and D) were at the baseline value of 600 ms. Short and Long events were always presented at their baseline frequency (i.e., 12).

Event frequency. For the event type manipulated across conditions (A or B), the frequency of that event type was manipulated by a factor of four relative to the baseline. The Few conditions received 3 presentations/dyad of A (or B) events. The Many conditions received 48 presentations/dyad of A (or B) events. The frequencies of all other events (B [or A] and C) were held at the baseline value of 12 per dyad, plus a total of 12 D events. Few and Many events were always presented at their baseline duration (i.e., 600 ms).

Statistical Analysis

Repeated-measures ANOVAs on relatedness ratings provided an omnibus analysis of the main effects and interactions among factors in the study. Specifically, a 3 (Dyad Number: 1 vs. 3 vs. 5) x 2 (Event Type: A vs. B) x 2 (Frequency: Few vs. Many) ANOVA was used to evaluate whether the effect of frequency depended on the number of dyads, event type, or both. Similarly, a 3 (Dyad Number) x 2 (Event Type) x 2 (Duration: Short vs. Long) ANOVA was used to

evaluate whether the effect of duration depended on the number of dyads or type of event. Notice that these ANOVAs omitted the baseline condition to avoid repeated use across analyses of the data from this condition. In principle, the baseline condition could have been repeated for each of the frequency and duration factors, which would have permitted tests of these factors with three instead of two levels. However, this approach would have given outsized weight to the baseline conditions. In practice, we also conducted analyses in which the baseline scores were reused, and the results of this analysis closely matched the results reported below. Notice that other analyses retained the baseline conditions in a way that permitted tests of three-level factors.

All p-values for repeated-measures ANOVAs were based on the Greenhouse-Geisser correction for nonsphericity. Effect sizes and corrections for nonsphericity were computed using the Psych Report (Mackenzie & Dudschig, 2022) and EZ (Lawrence, 2016) libraries in R (R Core Team, 2023). In addition, for each of the A and B event types, a linear mixed model (Kuznetsova et al., 2017) was used to evaluate specific hypotheses about the relationships between frequency (*f*) and number of dyads as independent variables and relatedness ratings (rating) as a criterion variable.

$$rating = \beta_1(f) + \beta_2(dyads) + \beta_3(f{:}dyads) + subject + error$$

A similar model was used to evaluate the effect of duration.

$$rating = \beta_1(d) + \beta_2(dyads) + \beta_3(d{:}dyads) + subject + error$$

where *f*, *d*, and *dyads* were the numerical frequencies, durations, and dyads in a stream, respectively. Proportional contrasts were -6, -3, and 9 for frequencies of 3, 12, and 48 and durations of 150, 600, and 2400 ms, respectively. Number of dyads was recoded such that the 1 dyad condition was assigned a value of -1, the 3 dyad condition was assigned a value of 0, and

the 5 dyad condition was assigned a value of +1. The coding of variables had no appreciable effect on the results of the analysis. Similar results were obtained with simple scaling and centering of variables. A stepwise algorithm for selection of predictors was used to identify the most parsimonious model. Repeated-measures ANOVAs require list-wise deletion of incomplete data, where all data from a participant is deleted if any of their responses are missing. To confirm that our results were not dependent on listwise deletion, we conducted a linear mixed model analysis based on pairwise deletion of missing data (i.e., removing only those conditions for which test ratings failed to meet the inclusion criteria described above). Thus, ANOVAs were based on data from 31 participants, and the linear regression analysis was based on data from all 59 participants. In all experiments, Bayesian tests of the null hypothesis were applied in cases where the null remained plausible and conventional significance tests did not reveal an effect. Bayes factors were computed using listwise deletion of missing data, which was analyzed using the lmBF() function in the BayesFactor library (Morey and Rouder, 2022). Figures were created using pairwise deletion of missing data.

Results

The results of Experiment 1 are summarized in Figure 2. The first column of the figure shows the effects of duration and frequency for A and B events in the One Dyad condition. The results for frequency in the One Dyad condition are like those of Murphy et al. (2022). A large effect of A frequency was observed, a smaller effect of B frequency was observed, and duration did not affect relatedness ratings. The top two rows of Figure 2 show the effect of frequency (A events = row 1, B events = row 2) on relatedness ratings. Clearly, increases in the frequency of A events resulted in increases in relatedness ratings. The relationship between the frequency of B events and relatedness ratings is less obvious. Across all levels of dyad number, a tendency

towards a decrease in relatedness ratings was seen with increases in the frequency of B events. However, this tendency appears to be unreliable as virtually no relationship exists among the Three Dyad conditions. The bottom two rows of Figure 2 show the relationship between event duration and relatedness ratings for A events (row 3) and B events (row 4). The pattern of results in the One Dyad conditions for Duration is like that reported by Murphy et al. (2022), who observed no effect of event duration on relatedness ratings. These impressions were supported by inferential statistics.

Omnibus test of frequency effects. A 3 (Dyad Number: 1 vs. 3 vs. 5) x 2 (Event Type: A vs. B) x 2 (Frequency: Few vs. Many) repeated-measures ANOVA on relatedness ratings that ignored the baseline condition detected a main effect of event frequency, $F(1, 30) = 12.79$, $p = .001$, $\eta_p^2 = .30$, which suggests that increases in the number of events produced increases in relatedness ratings. The analysis also detected a main effect of number of dyads, $F(2, 60) = 6.46$, Greenhouse-Geisser $p = .006$, $\eta_p^2 = .18$, indicating that relatedness ratings decreased as a function of the number of dyads. The effect of event type (A vs. B) was significant, $F(1, 30) = 5.02$, $p = .03$, $\eta_p^2 = .14$. Most importantly, the analysis detected an interaction between event type and frequency, $F(1, 30) = 18.56$, $p = .0002$, $\eta_p^2 = .38$, which suggests that changes in A event frequency and B event frequency produced different effects on relatedness ratings. Neither the interaction between frequency and number of dyads, $F(2, 60) = 0.63$, Greenhouse-Geisser $p = .51$, nor the interaction between event type and number of dyads, $F(2, 60) = 0.59$, Greenhouse-Geisser $p = .54$, was significant. The three-way interaction was also nonsignificant, $F(2, 116) = 0.02$, Greenhouse-Geisser $p = .93$. The source of the interaction and other specific hypotheses were evaluated by testing the linear mixed models described above.

LMM analysis of the effect of A frequency. A test of a linear mixed model of relatedness ratings in the conditions that varied the number of dyads and the frequency of A events with 600-ms A events detected a significant coefficient for the frequency of A events, β = 0.25, 95% CI [0.19, 0.31], $t(429.45) = 8.01$, $p < .0001$, and a significant coefficient for the number of dyads, β = -0.78, 95% CI [-1.27, -0.29], $t(431.34) = -3.13$, $p = .002$. The coefficient for the interaction was not significantly different from zero, β = -0.008, 95% CI[-0.08, 0.07], $t(432.33) = -0.22$, $p = .82$. Thus, increases in the frequency of A events produced increases in ratings of the relatedness of the cue and outcome, and increases in the number of dyads produced decreases in ratings. A backward stepwise algorithm for model selection favored a reduced model that omitted the interaction term. Specifically, a comparison between a model that included the interaction term and a model that omitted the interaction term detected evidence in support of the null hypothesis, $BF_{10} = 0.18$. Thus, increases in the frequency of A events affected relatedness ratings in a way that did not depend on the number of dyads. Also, increases in the number of dyads resulted in reduced relatedness ratings, independent of the frequency of A events.

LMM analysis of the effect of B frequency. An analysis of conditions that varied the number of dyads and frequency of B events (all events 600 ms) detected a significant frequency coefficient, β = -0.09, 95% CI [-0.15, -0.03], $t(421.03) = -2.92$, $p = .004$, and a significant dyad number coefficient, β = -0.52, 95% CI [-1.01, -0.05], $t(423.28) = -2.18$, $p = .03$. The interaction between dyad number and frequency was nonsignificant, β = -0.01, 95% CI [-0.08, 0.06], $t(420.37) = -0.26$, $p = .79$. A model selection and comparison procedure for B events produced results similar to the results obtained for A events. Specifically, a reduced model that omitted the interaction between dyad and frequency was identified. A comparison between it and a model that included the interaction term revealed support for the null hypothesis concerning the

interaction, $BF_{10} = 0.27$. Thus, increases in the frequency of B events or increases in the number of dyads resulted in reductions in relatedness ratings in the conditions that used 600-ms events. The lack of an interaction between the frequency and dyad number indicates that the effect of frequency occurs across training with multiple dyads. The lack of an effect of B frequency in the 3-Dyad conditions (see Figure 2) seems to be an anomaly.

Omnibus analysis of the effect of duration. A 3 (Dyad Number: 1 vs. 3 vs. 5) x 2 (Event Type: A vs. B) x 2 (Duration: Short vs. Long) repeated-measures ANOVA that ignored the baseline condition found an effect of number of dyads, $F(2, 60) = 5.71$, Greenhouse-Geisser $p = .006$, $\eta_p^2 = .10$. This analysis also detected two-way interactions between duration and event type, $F(1, 30) = 7.08$, $p = .01$, $\eta_p^2 = .19$, and between event type and dyads, $F(2, 60) = 7.09$, Greenhouse-Geisser $p = .002$, $\eta_p^2 = .19$. Neither the interaction between duration and number of dyads, $F(2, 60) = 0.23$, Greenhouse-Geisser $p = .79$, nor the three-way interaction was significant, $F(2, 60) = 2.71$, Greenhouse-Geisser $p = .07$. Neither the main effect of event type, $F(1, 30) = 0.54$, $p = .47$, nor the main effect of duration, $F(1, 30) = 3.18$, $p = .08$, was significant. Whereas the main effect of duration was not significant, duration clearly impacted ratings through an interaction between duration and event type.

LMM analysis of the effect of A duration. A linear mixed model was used to analyze the conditions that varied the number of dyads and the duration of A events among conditions that presented the baseline number of A events. This analysis detected a significant effect of A duration, $\beta = 0.10$, 95% CI [0.04, 0.16], $t(434.51) = 3.51$, $p = .0005$. The coefficients for dyad number, $\beta = -0.43$, 95% CI [-0.89, 0.02], $t(434.71) = -1.88$, $p = .06$, and for the interaction between dyad number and duration, $\beta = .06$, 95% CI [-0.01, 0.13], $t(434.33) = 1.67$, $p = .10$, were not significantly different from zero. The failure to detect a significant effect of dyad

number and a significant interaction between dyad number and duration seems to stem from a lack of statistical sensitivity rather than evidence supporting the null hypothesis because a Bayesian analysis failed to find support for the null hypothesis. The model selection algorithm retained only the duration independent variable. A comparison between a model that included only duration and the full model failed to provide support for the null hypothesis, $BF_{10} = .35$. Similarly, a comparison between a model that omitted only the interaction between number of dyads and duration failed to provide support for the null hypothesis, $BF_{10} = .40$. Thus, duration of A events affected relatedness ratings. While the number of dyads did not have a significant effect on relatedness ratings, there was no support for the hypothesis that number of dyads had a true null effect.

LMM analysis of the effect of B duration. A model of the effect of B duration failed to detect an effect, β = -0.03, 95% CI[-0.09, 0.03], $t(428.80) = -0.95$, $p = .34$. A significant effect of dyad number was observed, β = -0.53, 95% CI[-1.00, -0.04], $t(429.20) = -2.14$, $p = .03$. The interaction between dyad number and B duration was not significant, β = -0.03, 95% CI[-0.10, 0.04], $t(428.22) = -0.78$, $p = .44$. Here, the lack of a significant effect of duration should not be interpreted as support for the null hypothesis. A model selection algorithm retained only the number of dyads as an independent variable. However, a comparison between the full model and a reduced model that included only number of dyads as a factor provided rather modest support for the null hypothesis, $BF_{10} = 0.31$. Thus, it is unclear to what extent B duration affected relatedness ratings.

Discussion

Increases in the number of dyads clearly result in a decrease in relatedness ratings. This is not surprising, as with more dyads, it should be more difficult for participants to learn dyad

membership. Additionally, the present experiment, like Murphy et al.'s (2022), detected effects of both A and B frequencies. However, across the different levels of dyad number, the present analysis found a significant effect of A duration. Moreover, none of the analyses above supported a null effect of duration. While the effect of A duration was not significant in the One Dyad conditions (see Figure 2), the effect was significant in the Five Dyads conditions. This suggests a boundary condition to the relative insensitivity to event duration observed by Murphy et al. with one dyad. With five dyads, duration of A appears to affect ratings of relatedness. Seemingly, the effect of duration increases with increases in dyad number. However, this is merely a possibility because neither the interaction between dyad number and A event duration nor between dyad number and B event duration was significant. The primary difference between training with multiple dyads and training with only one dyad is that the former seems to result in some sensitivity to event duration whereas the latter has consistently produced no effect of event duration. Experiment 2 further examined this possibility by testing whether the free lunch effect (i.e., adjusted frequency effect) can be obtained when training contains multiple dyads.

Experiment 2

Experiment 1 detected an effect of A event duration on ratings of cue-outcome relatedness. This suggests a boundary condition to Murphy et al.'s central observation that duration is less important than frequency in affecting contingency learning. Specifically, while event duration seems unimportant for learning one and potentially three cue-outcome dyads, duration appears to be important for learning about five cue-outcome dyads. All conditions in Experiment 2 used five dyads. The primary purpose of Experiment 2 was to directly compare the effects of frequency and duration on relatedness ratings by accompanying increases in frequency with proportional decreases in duration. That is, in Experiment 2, event duration was

manipulated inversely to event frequency such that changes in frequency equated total exposure to the event type across a condition. In addition, Experiment 2 sought to test whether the effect of duration seen in Experiment 1 would generalize to a different range of values. Durations of 1000, 3000, and 9000 ms were used in Experiment 2.

Using longer events in Experiment 2 necessitated a reduction in the number of conditions per participant so that participants could complete the task in a single 60-minute session. Thus, Experiment 2 was divided into two sub-experiments: Experiment 2a tested the effects of A-event frequency and duration, and Experiment 2b tested the effects of B-event frequency and duration. The design of Experiment 2 is summarized in Table 3. Each condition in these experiments altered relative to baseline the simple frequency of the target event, the duration of the event, or the adjusted frequency of the event. Adjusted frequency manipulations involved inversely changing the duration of an event relative to changes in the event frequency such that total exposure to the event was matched between high (i.e., Many) and low (i.e., Few) adjusted frequency conditions. The analysis included Exposure Level to describe both the number and duration of trials, with High (Many and Long) and Low (Few and Short). Thus, each sub-experiment contained a repeated-measures 2 (Exposure Level: High vs. Low) x 3 (Trial Variation: Frequency vs. Duration vs. Adjusted Frequency) fully factorial design.

Experiment 2a Methods

Participants

Experiment 2a sought to recruit $N = 59$ participants, as was done in Experiment 1. However, in Experiment 2, participants were recruited from the Binghamton University research participation pool instead of Mechanical Turk. A sample of university students was used instead

of a crowdsourced sample because this experiment was conducted during the academic year when university students were available to serve in the experiments. Only 49 participants began the experiment, and of these, only 42 varied their responses across conditions. That is, seven participants provided identical ratings to all conditions and consequently were excluded from all analyses, leaving $N = 42$. Participants included 28 females and 21 males with an average age of 19 years (SD = 1.27). Participants were awarded course credit for completion of the study. The task was completed online using the Gorilla programming language and platform, as in Experiment 1. Six additional participants entered some of their ratings with a delay of greater than 30 seconds. Those ratings were coded as missing data. A 'missing completely at random' analysis was computed as in Experiment 1. This analysis failed to detect a pattern in the conditions from which participants were excluded, $\chi^2(23) = 17.30, p = .79$.

Stimuli

The stimuli used in Experiment 2a were identical to those used in Experiment 1. However, as Experiment 2a included fewer conditions (6), only 30 unique cue–outcome dyads, rather than 81, were used.

Procedure

The procedure in Experiment 2a was similar to that of Experiment 1, with the following exceptions. Experiment 2a did not include the same baseline condition as E1, but all conditions included baseline frequencies and durations of B, C, and D events. The baseline frequency of events A, B, and C in Experiment 2a was 9 per dyad (45 total per condition), and the baseline duration was 3000 ms. Nine D events in total were presented in each condition.

*Frequency***.** In the Low Frequency condition (i.e., Few), participants received only 3 A events per dyad (15 total) in addition to the baseline frequencies of B, C, and D events. Experiment 2a used block randomization as was done in Experiment 1. The Low Frequency condition consisted of three blocks, each containing one A event and three each of the B and C events per dyad. Each block also contained three D events. Events were randomly interspersed within a block. The High Frequency condition (i.e., Many) was treated similarly except that the A events of each dyad were presented nine times per block. Thus, in the Low Frequency condition, a total of three A events per dyad were presented, and in the High Frequency condition, a total of 27 A events per dyad were presented. All event durations were 3000 ms in the Low and High Frequency conditions. Consequently, total exposure to A events was 9 seconds per dyad in the Low Frequency condition and 81 seconds per dyad in the High Frequency condition.

Duration. In the Low Duration (i.e., Short) and High Duration (i.e., Long) conditions, participants received the baseline frequencies of all events. Thus, nine of each of the A, B, and C events per dyad (45 total) were presented in addition to nine D events. The duration of A events was 1000 ms in the Low Duration condition and 9000 ms in the High Duration condition. All other event durations (i.e., B, C, and D) were 3000 ms. Block randomization matched what was done in the Frequency conditions. That is, each block contained three A, B, and C events per dyad and three D events, randomly interspersed within a block. Each of the A, B, and C events was presented a total of nine times per dyad. Thus, total exposure to A events was 9 seconds per dyad in the Low Duration condition and 81 seconds per dyad in the High Duration condition.

Adjusted Frequency. In the Low Adjusted Frequency condition, participants received three A events per dyad (one in each block), with each A event lasting for 9000 ms., and in the

High Adjusted Frequency condition, 27 A events per dyad (nine in each block), with each lasting 1000 ms. Thus, total exposure to A events was 27 seconds in both the Low Adjusted Frequency and High Adjusted Frequency conditions. Each of the B and C events was presented a total of nine times per dyad, and there were 9 D events. As in the other conditions, trials were equally distributed across three blocks with randomization of trial order within blocks.

Statistical analysis

A 2 (Exposure Level: Low vs. High) x 3 (Trial Variation: Frequency vs. Duration vs. Adjusted Frequency) repeated-measures ANOVA on ratings of relatedness was used to evaluate whether the frequency, duration, or adjusted frequency of A events affected ratings. Planned comparisons and standardized effect sizes computed with the emmeans library (Lenth, 2023) were used to test specific hypotheses. The estimate of σ was obtained by finding the standard deviation of the residuals. ANOVAs and planned comparisons were based on the 42 participants who provided complete data. Most of the benefits of using a linear mixed model that motivated its use in Experiment 1 were absent from Experiment 2a because exposure level was effectively categorical, using only two levels: Low and High. However, we were interested in testing whether our results depended on our exclusion criteria. Thus, we conducted an LMM analysis similar to that described in Experiment 1, except that it used only the categorical factors described for the ANOVA. The results of an LMM analysis based on the full 42 participants agreed with the results of the ANOVA based on 36 participants who provided full data. For expediency, we omit any description of these LMM analyses with 42 participants. A Bayesian analysis was used to identify the level of support for the null hypothesis in analyses that failed to reach the criterion for statistical significance.

Experiment 2a Results

The results of Experiment 2a are depicted in Figure 3, which reveals that all three A event variables affected relatedness ratings. In addition to the simple frequency and duration effects detected in Experiment 1, Experiment 2a detected an adjusted frequency effect such that increases in the frequency of A events resulted in increases in the relatedness ratings, even when increases in frequency were achieved by reducing the duration of each event.

The repeated-measures ANOVA described in the Methods section detected only a main effect of exposure level (High vs. Low), $F(1, 35) = 29.09$, $p < .0001$, $\eta_p^s = .45$. Neither the main effect of A event variation type (Frequency vs. Duration vs. Adjusted Frequency), $F(2, 70) = 0.37$, Greenhouse-Geisser $p = .69$, nor the interaction, $F(2, 70) = 1.49$, Greenhouse-Geisser $p = .23$, was significant. A Bayesian analysis compared a model that included only A event variation type and participant as factors to a baseline model that included only participant. This analysis detected strong support for the null hypothesis, $BF_{10} = 0.06$, indicating that there were no differences between high and low exposure across the event variation levels. Similarly, a comparison between the baseline model and a model that included only A event variation type and the interaction between event variation and exposure level revealed strong support for the null, $BF_{10} = .01$. Thus, the effect of increasing exposure level was similar across the frequency, duration, and adjusted frequency conditions.

Experiment 2b Methods

Participants

As in Experiment 2a, we aimed to recruit 59 participants. In Experiment 2b, we recruited 59 SUNY-Binghamton student participants, six of whom provided identical ratings to all conditions and were consequently excluded from all analyses. Among the remaining 53

participants, 10 delayed their responses more than 30 s on one or more test ratings. A 'missing completely at random' analysis failed to detect a pattern in the missing data, $\chi^2(29) = 15.25$, $p = 0.98$. As was done in Experiment 2a, the 10 participants who provided incomplete data were excluded from the analysis because the ANOVA used for categorical independent variables requires listwise deletion of missing data. Because of an error in record-keeping, the demographic data for participants were lost; however, given the same population was used as in Experiment 2a, age and gender distributions were likely highly similar. Participants who served in Experiment 2a were excluded from participating in Experiment 2b.

Stimuli, procedure, and statistical analysis

The methods used in Experiment 2b were identical to those used in Experiment 2a, except that B events were manipulated instead of A events. Thus, the baseline frequency (nine for each of the five dyads) and duration (3000 ms) were used for all A events in Experiment 2b. The frequency and duration of B events were manipulated relative to baseline, as was done with A events in Experiment 2a.

Experiment 2b Results

Experiment 2b results are summarized in Figure 4. Based on the figure, it is evident that neither frequency nor adjusted frequency affected relatedness ratings. These impressions were supported by a 3 (B event variation: Frequency vs. Duration vs. Adjusted Frequency) x 2 (High vs. Low) repeated-measures ANOVA, which failed to detect any main effects or interactions. Specifically, the effect of exposure level was nonsignificant, $F(1, 42) = 1.65$, $p = .21$, as was the effect of B event variation, $F(2, 84) = 0.38$, Greenhouse-Geisser $p = .38$, and the interaction, $F(2, 84) = 0.54$, Greenhouse-Geisser $p = .58$. A Bayesian model comparison between the full model

with both main effects and the interaction against a baseline model that contained only participant detected strong support for the null hypothesis, $BF_{10} = 0.004$. The difference between the high and low duration conditions was not significant, $t(125) = -1.59$, $p = .11$, which we mention because the bootstrapped confidence intervals in Figure 4 are nearly nonoverlapping across the condition means.

Experiment 2 Discussion

Experiment 2 revealed several effects. The simple, unadjusted frequency of A events affected ratings of cue-outcome relatedness at test, with greater frequencies corresponding to higher ratings of relatedness. Moreover, there was no effect of the simple frequency of B events in Experiment 2b. We suspect that Experiment 2b failed to achieve statistical significance because participants are generally less sensitive to the amount of exposure to B events than A events (Wasserman et al., 1990). This contrasts with the results of Murphy et al. (2022), who observed little or no effect of the duration of A, B, C, and D events with a single dyad in each stream of trails across many experiments (see also Castiello et al., 2022, and Witnauer et al., 2023). Lastly, there was an effect of A adjusted frequency in Experiment 2a. Although both the frequency and duration of A events affected relatedness ratings, it seems that the effect of frequency is stronger than the effect of duration. Given equivalent amounts of total exposure to A events, a higher frequency of A events resulted in greater ratings than when the same total exposure was divided into a smaller number of A events. Thus, there is a benefit to presenting shorter, more numerous A events. This adjusted frequency effect was not observed in Experiment 2b, which could be explained by our lack of sensitivity to the effect of either simple frequency or duration of B events. However, Experiment 2 did not test the generality of the adjusted frequency effect to response measures other than relatedness ratings.

Experiment 3

Experiment 1 documented that there is no effect of event duration when a stream of events contains only one cue-outcome dyad, but there is an effect of event duration when a stream contains five cue-outcome dyads. Experiment 2, using five dyads, showed that an effect of adjusted frequency (i.e., a free lunch) for A manipulations (but not B manipulations) occurs when a stream contains five dyads. Both Experiments 1 and 2 used contingency ratings (i.e., relatedness) as dependent variables. Experiment 3 used a procedure like that of Experiment 2a – namely, a procedure that could replicate the adjusted frequency effect - except for the following changes, the most important of which was the use of a cued-recall memory test instead of relatedness ratings as the dependent variable.

Instead of using five dyads during training, Experiment 3 used ten dyads, with all cues being tested in a randomly selected order after the last trial in each training stream, which presumably allowed us to obtain a more reliable memory score for the number of correctly recalled items for each condition by having 11 possible values (0 through 10, inclusive). Mean correct per condition with five dyads would have constituted a more limited scale, and preliminary data suggested that ceiling 'correct' means were likely with only five dyads. Previous experiments (e.g., Murphy et al., 2022) using the streaming procedure had used a warmup condition when contingency judgments were recorded. No warmup condition was included in Experiments 1 and 2 because those previous experiments failed to find any difference between the warmup condition and the corresponding experimental condition that used the same cue-outcome contingency. We used a warmup condition in Experiment 3 because this was our first attempt at measuring recall memory. Experiment 3 included a baseline condition like that used in Experiment 1. Increases in the number of dyads and experimental conditions necessitated

using a baseline duration (560 ms) slightly shorter than the baseline duration in Experiments 1 and 2 (i.e., 600 ms). In addition, only A events were included in the experiment. That is, all conditions omitted B, C, and D events because presentation of B and C events would be inconsistent with our efforts to model the acquisition of vocabulary in a foreign language. Overt D events were omitted because we wanted to use a procedure with the most direct correspondence to applications like learning a foreign vocabulary. Moreover, covert D events for each dyad were created by presentations of other dyads. For instance, presentations of dyads 2a-2b, 3a-3b, 3a-4b, etc., were similar to D events (i.e., ITIs) for dyad 1a-1b in that they all lacked both 1a and 1b, but contained more information than would be expected during a standard ITI. Lastly, instead of using dyads consisting of characters (cues) and drawings of easily named objects (outcomes), Experiment 3 used dyads consisting of pseudo-Swahili words (cues) and English words (outcomes). We selected new stimuli for Experiment 3 because recall for words could be more readily measured than visual objects (Experiments 1 and 2). Words were used to mimic a form of language learning in which people learn to translate (i.e., respond with the corresponding English) word-to-word pairs that we assumed participants had no prior knowledge of. The design of Experiment 3 is summarized in Table 4.

Methods

Participants

Sixty-four participants were recruited from Binghamton University's participant pool. Recruitment and compensation in Experiment 3 matched Experiment 2. In Experiment 3, the mean age was 19 years (SD = 0.88) with 44 females and 20 males. No participants needed to be excluded due to a lack of variation in contingency ratings across conditions. All ten cue-outcome dyads in each condition were assessed in each cued-recall test. We retained from previous

experiments the requirement that participants respond to all test questions in a condition in less than 30 seconds for their data for that condition to be retained. This resulted in the removal of at least one condition for 20 participants. A 'missing completely at random' test failed to identify a pattern in the missing scores, $\chi^2(93) = 110.89, p = .10$.

Stimuli

The word list consisted of 80 cue-outcome dyads. Only nouns and verbs were included in the word list. A total of 45 nouns and 45 verbs were used. The words were restricted to a maximum of two syllables in both languages. The list was divided into clusters of 10 dyads held constant across conditions (i.e., one cluster per condition) to ensure that words presented in a condition were not like each other, with clusters randomly assigned to conditions anew for each subject.

Swahili to English translation was used so subjects would respond with English words to minimize ambiguities arising from misspellings. Instead of using proper Swahili, we adapted the language and used artificial Swahili words (pseudo-Swahili) so that the words would fit our criteria. Actual Swahili words were used without modification if the direct translation of the Swahili word met the 1- or 2-syllable criteria. If the direct translation exceeded the 1- or 2-syllable criteria in Swahili, syllables inside the word were chosen that met the stated criteria.

The chosen words were analyzed for word frequency in everyday English usage and reading level (i.e., mean age of first acquisition). The mean frequency of occurrence for all 80 words was 391.23 per 1,000,000 words. The range of this frequency was from 1 - 1947.27. These frequency numbers were generated using SUBTLEX, which compiles American subtitles to

generate a frequency system (Brysbaert & New, 2009). The average age of acquisition of these words was 5.33 years old, with a range of 2.37 - 10.33 (Kuperman et al., 2012).

Procedure

Participants were instructed to recall the missing member of each dyad (see Appendix B for instructions). In addition to the experimental conditions, a warmup condition was presented to participants before the experimental conditions. The warmup condition included 12 A events at 560 ms for each of the ten dyads. These 120 total trials were presented in 3 blocks, with each block containing four A events for each dyad, for a total of 40 A events/block. Within each block, the order of trials was randomly selected. The 'Warmup' contingency was identical to the Baseline condition except that it used different cue-outcome dyads and did not occur at a randomly determined position in the sequence of conditions.

Testing of memory occurred immediately after the last event in each condition. Each test was accompanied by text instructing participants.

Please translate the foreign word into English followed by <Enter>. If you cannot remember, guess. Press <Spacebar> to continue.

This was repeated ten times, once for each foreign word in that condition.

Statistical analysis

The statistical analysis for Experiment 3 was similar to that of Experiment 1, except that the number of correctly recalled items at test was the dependent variable. Responses to individual cued-recall questions were coded as correct or incorrect using the approximate string-

matching algorithm provided by the agrep function in R (R Core Team, 2023). The maximum distance between the participant's answer and the correct answer was set to .35, which corresponds to requiring 65% of the characters in the transformation to match. This value was chosen so that a three-letter word would be coded as a correct answer if one of the letters was incorrect. All of the analyses below were unaffected by using a lower value (.10). This algorithm was case-insensitive. In addition to the automated coding described above, manual coding with some tolerance for typos was done. The results based on manual coding closely approximated the results based on automated coding. We report only the results of automated coding for reproducibility purposes.

A 3 (Trial variation: Frequency vs. Duration vs. Adjusted frequency) x 2 (Exposure: High vs. Low) repeated-measures ANOVA on the number of correctly recalled memory items, omitting the baseline condition, was used as an omnibus analysis. Omission of the baseline condition was necessary because the experimental design included only one baseline condition as a level of exposure for all three levels of trial variation. In principle, it would be possible to repeat the baseline condition three times and add a third level to the frequency factor (or the duration factor). We think that this approach would be methodologically problematic because it would give the single baseline condition three times the impact on the analysis than any other condition. Thus, we do not report the results of an omnibus analysis that repeated the baseline condition (although it agreed with the results reported below). The ANOVA was based on listwise deletion of missing data. Separate linear mixed models using pairwise deletion of missing data were used to analyze the individual effects of duration and frequency. A linear mixed model of memory scores contained either frequency (numeric coding) and level of adjustment (categorical coding: adjusted = 1, unadjusted = -1) or duration as factors. The model

of frequency was applied to scores from frequency and adjusted frequency conditions, with the scores from the baseline condition duplicated.

$$correct = \beta_1(f) + \beta_2(adjustment) + \beta_3(f{:}adjustment) + subject + error$$

Adjustment was coded such that -1 = no adjustment and +1 = adjustment. The model used to evaluate the effect of duration was similar, except that it did not consider adjustment as a factor.

$$correct = \beta_1(d) + subject + error$$

Duplication of the baseline condition did not appreciably impact the results. Importantly, these statistical models were applied only to scores in their respective conditions. That is, the analysis of the frequency model did not consider scores in the short and long conditions (see Table 4), and the analysis of the duration model did not consider scores in the Few, Many, Few Adjusted, and Many Adjusted conditions. This was done to equate the number of scores in each condition. The coding of independent variables used in Experiment 1 was used again in Experiment 3 to facilitate comparisons between regression coefficients, even though it was not required by the duration model that included only one numeric independent variable. Proportional contrasts were -6, -3, and 9 for frequencies of 3, 12, and 48 and durations of 140, 560, and 2240 ms, respectively.

Results and Discussion

The left panel of Figure 5 summarizes the results of Experiment 3 in the frequency and adjusted frequency conditions, and the right panel summarizes the results in the duration conditions. There was an effect of simple frequency, with increases in the number of A events leading to an increase in the number of correctly recalled items at test. In addition, no free lunch was observed as increases in the duration of A events resulted in strong increases in recall at test.

In fact, increases in adjusted frequency decreased the number of outcomes recalled, presumably owing to the shorter duration of the trials used in conditions with higher adjusted frequency. This is consistent with observations in the duration conditions. Specifically, there was a strong effect of duration, with longer events giving rise to higher memory scores.

A 3 (A Event Trial Variation: Frequency vs. Duration vs. Adjusted Frequency) x 2 (Exposure: High vs. Low) repeated-measures ANOVA that omitted the baseline condition detected a main effect of exposure level, $F(1, 43) = 110.37$, $p < .0001$, $\eta_p^2 = .72$, a main effect of trial variation, $F(2, 86) = 3.95$, Greenhouse-Geisser $p = .02$, $\eta_p^2 = .08$, and an interaction, $F(2, 86) = 93.16$, Greenhouse-Geisser $p < .0001$, $\eta_p^2 = .68$. Importantly, the effect of increasing exposure level depended on whether the increase in exposure was achieved by an increase in frequency, duration, or adjusted frequency. A linear mixed model of the effect of frequency on memory scores detected no effect of frequency, β = 0.006, 95% CI[-0.03, 0.04], $t(285.79) = 0.36$, $p = .72$, and no effect of adjustment, β = -0.0004, 95% CI[-0.20, 0.20], $t(284.87) = -0.003$, $p = .997$. The coefficient for the interaction between frequency and adjustment was significantly different from zero, β = -0.21, 95% CI[-0.24, -0.18], $t(286.13) = -13.08$, $p < .0001$. Thus, the effect of frequency depended on whether the duration of the event was modified inversely to the increase in frequency so that the total amount of exposure to the event was unchanged. Notice that regression equations for either simple frequency or adjusted frequency had significant positive and negative slopes, respectively (see Figure 5). Thus, increases in simple frequency resulted in increases in the number of correctly recalled items, whereas increases in adjusted frequency resulted in decreases in recall. A model of the effect of duration on memory scores in the duration conditions detected a significant coefficient for event duration, β = 0.37, 95%

CI[0.31, 0.42], $t(120.00) = 13.09$, $p < .0001$. Thus, increases in the duration of A events produced improvements in memory scores.

Increases in either the frequency or duration of A events produced increases in memory scores. Of course, both of those manipulations were confounded with changes in the total amount of exposure to A events. The total times that participants were exposed to the Many and Long conditions were greater than in the Baseline condition, which was greater than the Few or Short conditions. In the adjusted conditions, increases in A event frequency produced decreases in A event duration such that all conditions used the same total amount of exposure to A events. The results of Experiment 3 failed to replicate the observation of a free lunch. Instead, they suggest that event duration is more important than event frequency in determining memory scores, which is a limitation to the critical observations of Murphy et al. (2022). In Experiment 3, the free lunch was abolished, presumably because training consisted of ten cue-outcome dyads, although we cannot preclude the use of different stimuli (pseudo-Swahili) as a factor. The results of Experiment 3 using a memory measurement agree with the results of Experiment 1 using Likert ratings in suggesting that increases in the number of dyads resulted in a decrease in the free lunch effect. The similarity between Experiments 1 and 3 indicates a correspondence between the Likert ratings used in most streaming experiments and a direct measurement of memory, such as cued recall.

Experiment 4

Experiment 4 tested whether contingency ratings agree with cued-recall scores using a procedure similar to that of Experiment 3, except that it included a measurement of contingency ratings. The design of Experiment 4 is summarized in Table 5. Experiment 4 included only one kind of test after each condition, either a contingency rating or a cued-recall test. The inclusion

of test type as a factor in Experiment 4 necessitated the following changes to the design of Experiment 4 relative to Experiment 3 to have sessions of feasible length: a) elimination of the baseline condition, b) manipulation of frequency, duration, or adjusted frequency by a factor of three instead of by a factor of four, and c) inclusion of only five dyads per condition instead of ten. Experiment 4 used a fully factorial and within-subjects design with test type (contingency vs. memory), trial variation (frequency vs. duration vs. adjusted frequency), and exposure level (high vs. low) as factors. All conditions included nine B and nine C cell events per dyad, plus a total of nine D events. These additional events were all 1000 ms in duration. No warmup condition was included in the experiment because, in Experiment 3, the warmup condition did not differ from the baseline condition. As in Experiments 1 through 3, all cue-outcome pairings (A events) were presented with a simultaneous cue and outcome.

Methods

Participants

Experiment 4 used both ratings of relatedness and a memory recall score. Based on power analyses adjusted for time in the academic term (statistical noise in our data from undergraduate participants increases later in each term), we decided to sample a larger number of participants from SUNY Binghamton's participant pool in Experiment 4 than were used in Experiments 1 through 3. Hence, we sought $N = 110$ participants, but only 90 participants signed up. Participants consisted of 73 females, 16 males, and one nonbinary student, with a mean age of 19 years ($SD = 0.81$). A *missing completely at random* test in R (Yanagida, 2023) was used to test whether there was a pattern in the missing data. This test failed to reach significance, $\chi^2(245) = 274.87, p = .09$. Only 59 participants provided data for all conditions. Once again, we used listwise deletion of missing data for statistics based on analysis of variance.

Procedure

Experiment 4 used a method like that of Experiment 3 except for the following changes. Twelve clusters of five English-Pseudo-Swahili dyads served as cues and outcomes. B and C events were included in each stream. Nine 1000-ms presentations of each event type (B and C) for each dyad were used. In addition, nine D events were included in each stream.

Statistical analysis

A factorial, repeated-measures, multivariate analysis of variance (MANOVA) with memory score and relatedness rating as dependent variables was used as an omnibus test of main effects and interactions. A multivariate analysis was used because Experiment 4 included two different measurements: cued recall scores and relatedness ratings on a Likert scale. Univariate ANOVAs were used to test the main effects and interactions for each of the individual dependent variables. MANOVAs and ANOVAs used only those participants who provided complete data (i.e., participants were deleted listwise for exceeding 30 seconds on any test trial), ensuring equal numbers of observations within each condition. Planned comparisons based on the ANOVA model were used to test for differences between high and low levels of exposure within each trial variation level and kind of measurement. Because the baseline condition was omitted in the design, frequency, duration, and adjusted frequency factors contained only two levels. Thus, the LMM analysis was omitted because none of the hypotheses in this experiment were concerned with linear relationships between exposure and memory or relatedness ratings.

Results and Discussion

Figure 6 depicts the results of Experiment 4. Notice that Experiment 4 replicated the central results of Experiments 1 through 3. The left panel illustrates that an increase in

contingency ratings was observed when either the duration or the simple frequency of an event was increased. Moreover, there was no effect of adjusted frequency on these ratings. Importantly, there was strong agreement between the ordinal differences observed with contingency tests and memory tests. That is, the same pattern of condition means was observed with both contingency and memory tests. This adds to the generality of the results previously observed with the streaming procedure. The following inferential statistics confirmed these impressions.

The Pearson correlation coefficient relating memory scores to contingency ratings among the 59 participants who completed all conditions was significantly different from zero, $r = .23$, 95% CI[.13, .32], $t(352) = 4.62$, $p < .0001$. This indicates that participants showed high memory scores in the same conditions that elicited high contingency ratings. Notice that the strength of this correlation was likely reduced by participants' rating and remembering different items at different times. That is, ratings and memory scores were obtained in different tests and for different cue-outcome dyads, which should increase error variance in the analysis. A 3 (Trial Variation: Adjusted Frequency vs. Duration vs. Frequency) x 2 (Exposure Level: High vs. Low) repeated measures MANOVA was used as an omnibus analysis with memory scores and contingency ratings as dependent variables detected an effect of trial variation, Wald's λ (4) = 23.02, $p < .001$, an effect of exposure level, Wald's λ (2) = 31.98, $p < .001$, and an interaction between trial variation and exposure level, Wald's λ (4) = 34.23, $p < .001$. The interaction between trial variation and exposure level indicates that the effect of increasing exposure to A events depended on whether the increase in exposure involved an increase in frequency, duration, or adjusted frequency. The interaction was driven by the increase in both ratings and recall caused by increases in frequency or duration, but not adjusted frequency.

Memory scores. A univariate 3 (Trial Variation) x 2 (Exposure Level) repeated-measures ANOVA on memory scores from the 59 participants who provided complete data detected an effect of exposure, $F(1, 58) = 21.23$, $p < .0001$, $\eta_p^2 = .27$, indicating that greater exposure corresponds to higher contingency ratings across trial variations. An effect of trial variation was detected, $F(2, 116) = 5.44$, Greenhouse-Geisser corrected $p = .008$, $\eta_p^2 = .09$. An interaction between trial variation and exposure level showed that the effect of increasing exposure depended on whether that increase in exposure was achieved by increasing adjusted frequency, duration, or simple frequency, $F(2, 116) = 13.74$, Greenhouse-Geisser corrected $p < .0001$, $\eta_p^2 = .19$. Planned comparisons detected a difference between high and low simple frequencies, $t(168) = 2.75$, $p = .007$, $d = 0.48$ [0.13, 0.82]. Thus, the number of correctly recalled items increased with greater numbers of A events. Similarly, greater duration of A events resulted in more correctly recalled items, $t(168) = 5.98$, $p < .0001$, $d = 1.04$ [0.67, 1.40]. The difference between high and low adjusted frequency levels was nonsignificant, $t(168) = -1.86$, $p = .06$, $d = -0.32$ [-0.67, 0.02]. Thus, the results of the repeated-measures ANOVA on memory scores broadly agreed with the results of the MANOVA.

Relatedness ratings. The analysis of relatedness ratings was similar to the analysis of memory scores described above. A factorial repeated-measures ANOVA on relatedness ratings showed an effect of exposure, $F(1, 58) = 16.72$, $p = .0001$, $\eta^2_{generalized} = .04$, and an interaction between exposure level and trial variation, $F(2, 116) = 8.83$, Greenhouse-Geisser $p = .0003$, $\eta^2_{generalized} = .03$. The main effect of trial variation was not significant, $F(2, 116) = 1.63$, $p = .20$, $\eta^2_{generalized} = .008$. Notice that the effect of trial variation was not significant in the ANOVA on contingency ratings but was significant in the analysis of memory scores. Planned comparisons detected differences between high and low exposure levels in the frequency

condition, $t(173) = 2.75$, $p = .007$, $d = 0.52$, 95% CI of d [0.14, 0.89], and in the duration condition, $t(173) = 5.21$, $p < .0001$, $d = 0.98$, 95% CI of d [0.59, 1.37]. Thus, increases in either frequency or duration produced increases in relatedness ratings, with duration producing a slightly larger effect than frequency. Once again, no difference between high and low adjusted frequencies was observed, $t(173) = -0.60$, $p = .55$, $d = -0.11$, 95% CI of d [-0.48, 0.26].

The results of Experiment 4 replicated the difficulty in obtaining a free lunch effect (i.e., an adjusted frequency effect) when participants learn about multiple dyads. In contrast to Murphy et al.'s (2022) observations with one dyad, the present experiment with five dyads per condition detected a strong effect of A-event duration on relatedness ratings. Moreover, Experiment 4 documented strong agreement between relatedness ratings and memory scores, adding to the generality of the present observations.

Experiment 5

Experiment 4 found that the adjusted frequency effect vanished when training and testing consisted of multiple dyads when A events were manipulated. The purpose of Experiment 5 was to determine if this attenuation of the free lunch effect with increasing frequency of training trials extended to D events. Specifically, Experiment 5 tested whether (a) relatedness ratings would be sensitive to the durations of D events when training and testing consisted of ten dyads and (b) whether there is agreement between relatedness ratings (i.e., contingency ratings) and recall scores in manipulations of D events. Hence, Experiment 5 was similar in design to Experiment 4, except that it included a baseline condition and manipulated D events rather than A events. Table 6 summarizes the design of Experiment 5. Based on the results of Experiments 1 through 4, increases in both relatedness ratings and numbers of correctly recalled items might be expected from increases in either simple frequency or duration. Little or no effect of D event adjusted

frequency was expected. Thus, Experiment 5 was a 2 (Test Type: Memory vs. Relatedness Rating) x 3 (Trial Variation: Frequency vs. Duration vs. Adjusted Frequency) x 2 (Exposure Level: High vs. Low) repeated-measures factorial design plus a baseline condition in which the durations of D events were 550 ms and the frequency was 60. Across conditions, manipulation of exposure level decreased (low) or increased (high) frequencies or durations of D events by a factor of five relative to the baseline condition.

Methods

Participants

Eighty-seven participants recruited from Prolific served in the experiment, including 30 females, 54 males, and 3 participants who declined to be categorized. A crowdsourced sample was used because this experiment was conducted outside of the academic year. The mean age was 33 years (*SD* = 7.8). Of the 87 participants who began the experiment, only 86 responded to at least one test and only 62 provided complete data. A *'missing completely at random'* test was not significant, $\chi^2(219) = 213.09, p = .60$.

Procedure

The stimuli and instructions were identical to those used in Experiments 3 through 4. Measurements of relatedness ratings and cued-recall matched Experiment 3. In the baseline conditions, six A events and three B events were presented for 700 ms each for each of the ten dyads. Interspersed among these A and B events were 60 D events at 550 ms. These 150 total trials were presented in a block randomized order such that each of the three blocks consisted of 2 (x 10 dyads) A-cell events (700 ms), 1 (x 10 cues) B-cell events (700 ms) for each dyad, and 20 D-cell events (550 ms), for a total of 50 trials per block. Within each block, the order of these

50 trials was randomly selected anew for each participant. On A events, the pseudo-Swahili word was presented alone for 150 ms immediately before the addition of the English word dyad member to the screen. The English and pseudo-Swahili words that made up a given dyad were presented together on the screen for 400 ms. Then the pseudo-Swahili word was removed, and the English word remained on the screen for another 150 ms. Thus, pairings of pseudo-Swahili words and English translations were sequential, in contrast to Experiments 1-4, which used simultaneous pairings. Relative to the baseline conditions, either the frequency, duration, or adjusted frequency of D events was manipulated. A events (6 for 700 ms) and B events (3 for 700 ms) were identical in frequency and duration across conditions. In the few D conditions, four-fifths of the baseline D events were eliminated. In the short D conditions, each of the 60 D events was one-fifth of the baseline duration. In the few adjusted D conditions, four-fifths of the baseline D events were eliminated *and* the remaining D events were longer in duration by a factor of five. The Many, Long, and Many adjusted conditions were treated similarly. Thus, total exposure to D events was equated across the Baseline, Few adjusted, and Many adjusted conditions (33 seconds per condition). Total exposure to D events was equivalent in the Long and Many conditions (165 seconds each) and in the Short and Few conditions (6.6 seconds each). After training, testing consisted of either a sequential cued-recall test consisting of all ten training dyads or a test of relatedness rating for one of the dyads used in the stream, randomly chosen.

Statistical analysis

As in Experiment 3, statistics based on ANOVAs used listwise deletion of missing data, whereas statistics based on mixed models used pairwise deletion of missing data. A repeated-measures, factorial MANOVA was used as an omnibus analysis of relatedness ratings and recall

scores as dependent variables and with trial variation and exposure level as categorical factors. In addition, separate univariate ANOVAs on each of the dependent variables were conducted to evaluate the effects of exposure level on each of the dependent variables. The baseline condition was omitted from the MANOVA and the separate univariate ANOVAs that were used to test specific hypotheses about relatedness ratings or cued-recall. The motivation for excluding the baseline condition in the MANOVA and ANOVA analyses is identical to that described in Experiment 3. However, like Experiment 3, Experiment 5 involved follow-up tests that evaluated the effects of individual variables (D frequency, D adjusted frequency, and D duration) within each dependent variable. Here, the baseline condition was used such each of these factors had three levels. Once again, the results of the analysis did not depend on how we treated the baseline condition.

Results and Discussion

The results of Experiment 5 are summarized in Figure 7. A 3 (Trial Variation: Frequency vs. Duration vs. Adjusted Frequency) x 2 (Exposure Level: High vs. Low) repeated-measures MANOVA with cued-recall and relatedness ratings as dependent variables detected an effect of exposure level, Walds $\lambda(2) = 49.97$, $p < .001$, indicating that increases in frequency of D events produced increases in both relatedness ratings and cued-recall scores. Moreover, the MANOVA detected an interaction, Walds $\lambda(4) = 23.35$, $p < .001$, which is consistent with there being an effect of frequency and duration but not adjusted frequency. The main effect of trial variation was not significant, $\lambda(4) = 4.49$, $p = .34$. The MANOVA results suggest that the exposure level effect was similar between relatedness ratings and cued-recall. The simple bivariate correlation between memory scores and contingency ratings was not significant, $r = .09$ [-.00, .19], $t(432) = 1.96$, $p < .051$.

Relatedness ratings. A factorial repeated-measures ANOVA with the same factors and levels as the omnibus MANOVA was used to analyze relatedness ratings by the 62 participants who provided complete data. Only a main effect of exposure level (i.e., D-event frequency) was obtained, $F(1, 61) = 5.13$, $p = .03$, $\eta_p^2 = .08$. The main effect of trial variation was not significant, $F(2, 122) = 0.08$, Greenhouse-Geisser $p = .92$, and the interaction between trial variation and exposure level was not significant, $F(2, 122 = 1.54$, Greenhouse-Geisser $p = .22$. A linear mixed model of relatedness ratings from all 86 participants was tested. This model included frequency and adjustment as fixed effects and participant as a random effect. The effect of frequency was not significant, $\beta = 0.05$ [-0.02, 0.11], $t(417.66) = 1.38$, $p = .17$, nor was the coefficient for adjustment, $\beta = 0.09$ [-0.34, 0.52], $t(417.67) = 0.39$, $p = .69$. However, the interaction was significant, $\beta = -0.08$ [-0.15, -0.01], $t(418.28) = -2.30$, $p = .02$, suggesting that the effect of an increase in frequency on relatedness ratings depended on whether a decrease in duration accompanied the increase in frequency. That is, there was an increase in ratings with increases in simple frequency but no increase in ratings with increases in adjusted frequency (see Figure 7). However, it should be noted that a Bayesian comparison among the three adjusted frequency conditions (including baseline and including only those participants who completed all conditions) detected moderate support for the null, $BF_{10} = 0.16$. The nonsignificant effect of adjusted frequency, then, seems to have been caused by a true null effect of adjusted D event frequency rather than an insensitivity to the effect. Similar results were obtained from a linear mixed model analysis that included duration as a factor. The duration coefficient was not different from zero, $\beta = 0.03$ [-0.04, 0.09], $t(419.37) = 0.75$, $p = .46$. A Bayesian model comparison that included event duration in the numerator but not in the denominator detected moderate support for the null hypothesis, $BF_{10} = 0.18$.

Memory scores. A repeated-measures ANOVA with the same factors and levels as the MANOVA was used to analyze cued-recall scores from the 62 participants who provided complete data. This analysis revealed a main effect of exposure level, $F(2, 61) = 43.33$, $p < .0001$, $\eta_p^2 = .42$, and an interaction between exposure level and trial variation, $F(2, 122) = 9.14$, Greenhouse-Geisser $p = .0003$, $\eta_p^2 = .13$. The main effect of trial variation was not significant, $F(2, 122) = 2.05$, Greenhouse-Geisser $p = .13$. Thus, increases in exposure resulted in increases in cued-recall to a degree that depended on whether frequency, duration, or adjusted frequency was manipulated. A linear mixed model of the effects of frequency and adjustment as well as their interaction on cued-recall was tested. The coefficient for frequency was significant, β = 0.07 [0.04, 0.10], $t(386.25) = 5.05$, $p < .0001$. Thus, increases in the frequency of D events resulted in improved cued-recall performance. Adjustment had no effect on memory scores, β = -0.14 [-0.31, 0.32], $t(385.49) = -1.60$, $p = .11$. However, adjustment interacted with frequency, β = -0.06 [-0.09, -0.03], $t(386.14) = -4.26$, $p < .0001$. This indicates that the increases in only the simple frequency of D events affected cued recall; increases in adjusted frequency did not (see Figure 7). This is consistent with the results of Experiments 1-4 in showing only a simple frequency but not an adjusted frequency effect. Relatedly, an analysis of a linear mixed model of the effect of duration on memory scores detected an increase in memory with increases in D event duration, β = 0.11 [0.06, 0.15], $t(157.23) = 4.97$, $p < .0001$. Thus, training with ten dyads resulted in learning that was sensitive to the duration of D events in a way that is inconsistent with the free lunch effect.

General Discussion

The present experiments extended and identified both generalizations and limitations of Murphy et al.'s (2022) observation of the free lunch effect. The present experiments replicated

the basic observation that increases in the frequency of an event result in increases in the effect of that event on relatedness ratings. Moreover, when training consisted of one dyad (Experiment 1), ratings were relatively insensitive to manipulations of event duration. Similarly, when training consisted of three dyads (Experiment 2), a free lunch effect was observed for manipulations of A events. That is, increases in the frequency of an event resulted in increases in the effect of that event on ratings, even when total exposure to that event is held constant across conditions. Importantly, the present experiments also showed generally strong agreement between relatedness ratings and memory scores, except in Experiment 5, where this relationship was nonsignificant.

No change in relatedness ratings resulted from changes in exposure to B events; however, it is well established that variation in A events has a greater impact on contingency ratings than equivalent variation in B or C events, and variation in D events has the smallest impact (Murphy et al., 2022; Wasserman et al., 1990). Critically, a direct effect of A duration emerged with increases in the number of dyads in a condition in Experiment 1. The effect of duration was examined more closely in Experiment 2. Experiment 2a detected effects of frequency, duration, and adjusted frequency of A events, whereas Experiment 2b detected no effect of B exposure level. The results of Experiments 1 and 2 collectively suggest that the duration of an event becomes more important with increases in the number of dyads in a condition. In Experiment 3, a cued-recall test was used to test the effects of levels of exposure to A events when streams contained ten dyads. In that experiment, strong effects of frequency and duration were observed, and, strikingly, increases in the adjusted frequency of A events resulted in *decreases* in cued-recall performance at test, indicating that the effect of duration had become greater than that of frequency. Experiment 4 used five dyads per stream and tested both cued-recall performance and

relatedness ratings. With only 5 dyads per condition, there was no effect of adjusted frequency, but we observed strong effects of duration and simple frequency on relatedness ratings and cued-recall. Thus, the upper limit for the free lunch effect on cued-recall for A events in our preparation seems to be less than 5 dyads. Experiment 5 extended this observation from A events to D events with ten dyads. In Experiment 5, a strong effect of duration was observed and the increase in ratings or cued-recall that resulted from increases in frequency was eliminated when a proportionate reduction in per-D-event duration accompanied the increase in frequency. The present experiments in aggregate suggest that only manipulations of exposure level that increase total exposure to the event result in increased learning about the event when multiple dyads are present during initial acquisition.

That increase in the number of dyads within a condition increased the effect of trial duration possibly reflects the greater attentional or load demands required to relate multiple cues with multiple outcomes than a single cue with a single outcome. This account suggests future experiments in which a distraction task running during training or more complex target stimuli might result in an effect of trial duration with a single cue-outcome dyad, and much longer trials might result in a free lunch effect even with five or ten cue-outcome dyads.

Previous experiments from our laboratories have been interpreted in the context of contingency theory. Specifically, we have found that Δp (Allan, 1980) provides a good description of many effects observed in streamed trial procedures. According to Δp, performance at test is related to the change in the probability of an outcome occurring that is signaled by a cue. Specifically, Δp is the difference between two conditional probabilities: the probability of the outcome given the cue minus the probability of the outcome in the absence of the cue (i.e., experimental context alone). A events increase the conditional probability of the outcome given

the cue, and B events decrease that conditional probability. Similarly, C events increase the probability of the outcome given the absence of the cue, and D events decrease that conditional probability. Conditional probabilities are often assumed to be based on the simple frequencies of events alone and independent of the duration of the events. The probability of the outcome given the cue is typically computed as the simple number of A events divided by the total number of times the cue was presented (A events plus B events) without consideration of the duration of A and B events. Alternatively, conditional probabilities could be based on total exposure time to the different events (e.g., the total amount of time spent in all A events divided by the total amount of time in all A and B exposures).

To the extent that the free lunch effect occurs when the adjusted frequency of A or B events is increased (e.g., Experiment 1, one dyad), Δp provides the best description of the data when it is based on frequencies rather than total exposure. However, the present experiments suggest that whether (adjusted) frequency or total exposure best predicts performance at test depends on the number of target dyads being trained and the specific cell being manipulated. Of course, adjusted frequency and duration are not mutually exclusive. In Experiment 1 with contingency ratings after streams with five dyads, there was an adjusted frequency effect and an effect of duration. There was a strong adjusted frequency effect for A cell events when the stream contained five dyads (Experiment 2a), but there was no adjusted frequency effect for B events (Experiment 2b). Moreover, total exposure to A events affected ratings (Experiment 3 and potentially Experiment 1). Thus, Experiments 1-4 provide only partial support (in conditions with only one or three dyads) for a version of Δp that considers only frequency and ignores duration.

Experiment 5 is even more difficult to interpret in the context of Δp. The results of Experiment 5 extended the results of Experiments 1-4 to D events, suggesting that both relatedness ratings and cued-recall scores are sensitive to both frequency and duration of D events. Notice that even a version of Δp that considers the total duration of exposure to events A through D fails to explain the results of the present experiments. Generally, Δp fails to predict any effect of D cell manipulations in Experiment 5. Specifically, D events affect Δp by altering the probability of the outcome given the absence of the cue. However, that probability should have always been zero because the outcome was never presented in the absence of the cue (i.e., there were no C events) unless one considers the end of a cue-outcome pairing a C event. Thus, Δp provides only an incomplete description of the present results.

An alternative interpretation of Experiment 5 considers the roles of the context as a companion cue that competes with the target cue (e.g., Durlach, 1980; Witnauer & Miller, 2012). According to this analysis, increases in the frequency of D events reduce the potential of the context (the trial-marking frame in our procedure) to compete with the cue because the context extinguishes during each D event. The simplistic and mechanistic approach of this explanation is appealing. More generally, one could use a wide range of associative models to explain the adjusted frequency effect (e.g., Mackintosh, 1975; Miller & Matzel, 1988; Rescorla & Wagner, 1972). However, this explanation fails to address the very large difference between the effects of B and D events in the present experiments. It is unclear why additional presentations of the trial marker alone (i.e., D events) would affect cued recall and contingency judgments, but additional presentations of the cue and trial marker without the outcome (i.e., B events) had no appreciable effect on contingency judgments. However, it should be noted that this apparent difference between B and D events is based on a between-experiment comparison.

The present results are consistent with previous research on the effect of list length on human memory. Specifically, we found a decrease in relatedness ratings with increases in the number of dyads in Experiments 1 and 2. This parallels the decrease in accuracy of free recall that results from increases in the number of items on a list of to-be-recalled items (e.g., Roberts, 1972). A between-experiment comparison points to a similar effect in the present cued-recall data. That is, there was a decrease in the proportion of items correctly recalled when ten dyads were used in training relative to when only 5 dyads were used (Experiment 3). Thus, increasing the number of dyads appears to reduce both perceived contingency and recall accuracy.

Parts of the present results are consistent with the observation that similar levels of behavioral control in Pavlovian procedures occur under common ratios of intertrial interval to interstimulus interval (Gibbon & Balsam, 1981). Specifically, Gibbon and Balsam observed across conditioning preparations that the strength of a Pavlovian conditioned response is related to the ratio of two intervals:

Cycle (C): the average time between successive presentations of the US

Trial (T): the average amount of time in a trial

(i.e., CS-present time between CS-US pairings)

Greater levels of responding are observed for larger C/T ratios, and similar levels of responding are observed under constant C/T ratios. In the present experiments, increases in the frequency or duration of A or D events, but not increases in the adjusted frequency of these events, should increase the C/T ratio. In contrast, increases in the frequency or duration of B and C* events should decrease the C/T ratio. Thus, a free lunch effect is not predicted based on the C/T ratio. However, the central observation in the present experiments was that there are limits to the free

lunch effect achieved by manipulating adjusted frequency. Possibly, the C/T ratio predicts the effects of frequency and duration within the boundaries of the free lunch effect.

In summary, the present results suggest that increases in performance on both contingency (i.e., 'relatedness') and recall tests of simple associative learning tasks, with either simultaneous (Experiments 1-4) or sequential (Experiment 5) cue-outcome presentations, can be achieved by increases in the frequency of exposure to either cue-outcome pairings or context-alone events during training. This somewhat unsurprising effect is consistent with a growing literature pointing to the importance of frequency in controlling learning and memory (Murphy et al., 2022). However, the present results also suggest a boundary for the benefit of increased frequency when duration is reduced proportionally (i.e., the free lunch effect). That is, increases in adjusted frequency of an event were ineffective when training consisted of more than three dyads. The present experiments do not allow us to explain why proportional changes in trial frequency and duration sometimes have different effects. However, attention is a likely factor because attention ordinarily wanes as trial duration increases, whereas the onset of each new trial is expected to refocus attention.

Limitations

The central aim of the present experiments was to identify boundary conditions for the effect of adjusted frequency on contingency ratings and cued recall. Because the project explored the effect of several variables (dyad number, response modality, and event type), some of the most important conclusions that can be gleaned from the present experiments are supported in part or in total only from between-experiment comparisons. For example, in Experiment 5, manipulations of D event adjusted frequency failed to produce a free lunch effect when training consisted of 10 dyads when either contingency ratings or memory scores were measured. It

remains unknown whether a free lunch effect would generalize to a smaller number of dyads or to a different target event type. The results of Experiments 1-4 suggest that it was not because these experiments used fewer dyads or a different event type. However, each of those experiments examined the effect of a different type of event (A or B) or tested only one response modality.

The results of the present experiments are important primarily for empirical and potentially applied, as opposed to theoretical, reasons. The mechanism(s) involved in the free lunch effect remain unclear. For example, models of learning that consider time might be able to explain aspects of the adjusted frequency effect. For example, longer presentations of an event might trigger habituation such that a multiplicative increase in duration is less effective than a multiplicative increase in frequency, even when the increase in frequency is accompanied by adjustment of duration. However, the present experiments were not designed to test this explanation of the free lunch effect, and possibly a different mechanism is responsible for the reduced effectiveness of long events relative to short events. In any case, a successful explanation of the free lunch effect should be able to explain the effect's boundaries (e.g., no free lunch with multiple dyads).

Relatedly, it is possible that increases in the number of dyads created a greater cognitive load that allowed for some attentional benefit to be achieved by increasing event duration. We identified two possibilities concerning the increased effect of duration with increasing number of dyads. First, it is possible that increasing the number of dyads results in a free lunch effect only at a greater time scale than was used in the present experiments (e.g., with 5 dyads, greater learning may occur with 96 events at 1 second each than 6 events at 16 seconds each). Second, alternatively, it is possible that increasing the number of dyads results in no free lunch at any

baseline duration (e.g., with 5 dyads, the amount of learning may be determined exclusively by the total exposure to an event). The present experiments do not differentiate between these possibilities. The present results merely document that the free lunch effect does not occur with multiple dyads when A- or D-event adjusted frequency is manipulated within the range of values examined and either contingency ratings or cued-recall scores are measured.

References

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147–149. https://doi.org/10.3758/BF03334492

Anwyl-Irvine, A. L., Massonnie, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Aue, W. R., Criss, A. H., & Fischetti, N. W. (2012). Associative information in memory: Evidence from cued recall. *Journal of Memory and Language*, *66*(1), 109–122. https://doi.org/10.1016/j.jml.2011.08.002

Avila, E., & Sadoski, M. (1996). Exploring new applications of the keyword method to acquire English vocabulary. *Language Learning*, *46*(3), 379–395.

Bower, G. H. (1962). An association model for response and training variables in paired-associate learning. *Psychological Review*, *69*(1), 34–53. https://doi.org/10.1037/h0039023

Brady, T. F., Robinson, M. M., Williams, J. R., & Wixted, J. T. (2023). Measuring memory is harder than you think: How to avoid problematic measurement practices in memory research. *Psychonomic Bulletin & Review*, *30*(2), 421–449. https://doi.org/10.3758/s13423-022-02179-w

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Castiello, S., Miller, R. R., Witnauer, J. E., Alcaide, D. M., Fung, E., Pitliya, R. J., Morrissey, D. K. C., & Murphy, R. A. (2022). Benefiting from trial spacing without the cost of prolonged training:

Frequency, not duration, of trials with absent stimuli enhances perceived contingency. *Journal of Experimental Psychology: General*, *151*(8), 1772–1792. https://doi.org/10.1037/xge0001166

Criss, A. H., & Shiffrin, R. M. (2004). Pairs do not suffer interference from other types of pairs or single items in associative recognition. *Memory & Cognition*, *32*(8), 1284–1297. https://doi.org/10.3758/BF03206319

Crump, M. J. C., Hannah, S. D., Allan, L. G., & Hord, L. K. (2007). Contingency judgements on the fly. *Quarterly Journal of Experimental Psychology*, *60*(6), 753–761. https://doi.org/10.1080/17470210701257685

Durlach, P. J. (1983). Effect of signaling intertrial unconditioned stimuli in autoshaping. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*(4), 374.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*(3), 227–247. https://doi.org/10.1037/0096-3445.117.3.227

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*, 978–990.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lawrence, M. A. (2016). *ez: Easy Analysis and Visualization of Factorial Experiments*. https://CRAN.R-project.org/package=ez

Lenth, R. V. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://CRAN.R-project.org/package=emmeans

Mackenzie, I. G., & Dudschig, C. (2022). *psychReport: Reproducible Reports in Psychology*. https://CRAN.R-project.org/package=psychReport

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

Miller, R. R., & Matzel, L. D. (1988). The comparator hypothesis: A response rule for the expression of associations. In *Psychology of learning and motivation* (Vol. 22, pp. 51–92). Elsevier.

Morey, R. D., & Rouder, J. N. (2022). *BayesFactor: Computation of Bayes Factors for Common Designs*. https://CRAN.R-project.org/package=BayesFactor

Murdock, B. B. (1968). Serial order effects in short-term memory. *Journal of Experimental Psychology*, *76*(4, Pt.2), 1–15. https://doi.org/10.1037/h0025694

Murdock Jr., B. B. (1960). The immediate retention of unrelated words. *Journal of Experimental Psychology*, *60*(4), 222–234. https://doi.org/10.1037/h0045145

Murphy, R. A., Witnauer, J. E., Castiello, S., Tsvetkov, A., Li, A., Alcaide, D. M., & Miller, R. R. (2022). More frequent, shorter trials enhance acquisition in a training session: There is a free lunch! *Journal of Experimental Psychology: General*, *151*(1), 41–64. https://doi.org/10.1037/xge0000910

Pavlov, I. P. (1927). Conditioned Reflexes: Oxford University Press. *London, UK [Google Scholar]*.

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II, A. H. Black & W. F. Prokasy, Eds.* (pp. 64–99). Appleton-Century-Crofts.

Roberts, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *Journal of Experimental Psychology*, *92*(3), 365.

Shanks, D. R. (1985). Forward and Backward Blocking in Human Contingency Judgement. *The Quarterly Journal of Experimental Psychology Section B*, *37*(1b), 1–21. https://doi.org/10.1080/14640748508402082

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin*, *133*(6), 1038–1066. https://doi.org/10.1037/0033-2909.133.6.1038

Wasserman, E. A., Dorner, W., & Kao, S. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 509.

Witnauer, J. E., Castiello, S., Fung, E., Pitliya, R. J., Murphy, R. A., & Miller, R. R. (2023). Determinants of extinction in a streamed trial procedure. *Quarterly Journal of Experimental Psychology*, *76*(5), 1155–1176. https://doi.org/10.1177/17470218221110827

Witnauer, J. E., & Miller, R. R. (2012). Associative status of the training context determines the effectiveness of compound extinction. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*(1), 52–65. https://doi.org/10.1037/a0026333

Yanagida, T. (2023). *misty: Miscellaneous functions for structural equation models and other applications* (Version 0.6.5). https://rdocumentation.org/packages/misty/versions/0.6.5

Table 1. ***Overview of experiments***

Exp	Cues	Outcomes	Response	Training events	Dyads	Manipulated variables
1	Non-Latin characters	Drawings of objects	Rating	A, B, C, D	1, 3, or 5	• Cell: Whether exposure to A or B events was different from baseline (12 at 600 ms) • Number of Dyads: 1, 3, or 5 dyads per stream • Simple frequency of A or B relative to baseline (3, 12, or 48 events), confounded with total duration of exposure within a stream (1.8, 7.2, or 28.8 seconds of A or B per dyad) • Duration of A or B relative to baseline (150, 600, or 2400 ms), confounded with total duration of exposure to an event within a stream (1.8, 7.2, or 28.8 seconds of A or B per dyad) • Frequency and duration were partially crossed with each other. Frequency x Duration combinations were fully crossed with dyad number and cell.
2a	Non-Latin characters	Drawings of objects	Rating	A, B, C, D	5	• Simple frequency of A (3 or 27), confounded with total duration of exposure within a stream • Adjusted frequency of A, confounding frequency with individual event duration but holding total exposure constant. • Duration of A
2b	Non-Latin characters	Drawings of objects	Rating	A, B, C, D	5	• Simple frequency of B, confounded with total duration of exposure within a stream. • Adjusted frequency of B, confounding frequency with individual event duration but holding total exposure constant. • Duration of B
3	Pseudo-Swahili words	English translation	Recall	A, B, C, D	10	• Simple frequency of A, confounded with total duration of exposure within a stream. • Adjusted frequency of A, confounding frequency with individual event duration but holding total exposure constant. • Duration of A
4	Pseud-Swahili words	English translation	Recall and Rating	A, B, C, D	5	• Simple frequency of A, confounded with total duration of exposure within a stream. • Adjusted frequency of A, confounding frequency with individual event duration but holding total exposure constant. • Duration of A
5	Pseud-Swahili words	English translation	Recall and Rating	A, B, D	10	• Simple frequency of D relative to baseline, confounded with total duration of exposure to D within a stream. • Adjusted frequency of D relative to baseline, confounding frequency with individual event duration but holding total exposure constant. • Duration of D relative to baseline, confounded with total exposure to D within a stream.

Note: Exp = experiment, A-D are cue-outcome, cue alone, outcome alone, and neither cue nor outcome events in a streamed trial procedure.

Table 2. ***Design of Experiment 1***

		Duration (ms) of A or B events		
		150	**600**	**2400**
Frequency of A or B events	**3**		**Few**: 3 per dyad at 600 ms each **Total exposure:** 1.8 sec/dyad	
	12	**Short:** 12 per dyad at 150 ms each **Total exposure:** 1.8 sec/dyad	**Baseline:** 12 per dyad at 600 ms each **Total exposure:** 7.2 sec/dyad	**Long:** 12 per dyad at 2400 ms each **Total exposure:** 28.8 sec/dyad
	48		**Many:** 48 per dyad at 600 ms each **Total exposure:** 28.8 sec/dyad	

Note: The baseline frequencies of A, B, and C events were 12 per dyad, and the baseline frequency for D events was 12 across all dyads. A common baseline condition was used for each level of the dyad factor (i.e., the baseline condition was not run separately for the A and B conditions). Each cell in Table 2 reports only the frequency and duration of the event (A or B) that was different from Baseline. All frequencies except D events are per dyad. Each of the five exposure levels in Table 2 occurred in each of three different levels of dyad number (one, three, or five). Frequency of exposures (3, 12, 48 per dyad) and Duration of each exposure (150, 600, 2400 ms) were partially crossed, yielding 5 of the 9 possible Frequency × Duration combinations. Each Frequency × Duration condition was fully crossed with Dyad number (1, 3, 5) and Cell manipulated (A, B).

Table 3. ***Design of Experiment 2***

		Duration (ms) of A or B events		
		1000	**3000**	**9000**
Frequency of A or B events	**3**		**Few**: 3 per dyad at 3000 ms each **Total exposure:** 9 sec/dyad	**Few Adjusted:** 3 per dyad at 9000 ms each **Total exposure:** 27 sec/dyad
	9	**Short:** 9 per dyad at 1000 ms each **Total exposure:** 9 sec/dyad		**Long:** 9 per dyad at 9000 ms each **Total exposure:** 81 sec/dyad
	27	**Many Adjusted:** 27 per dyad at 1000 ms each **Total exposure:** 27 sec/dyad	**Many:** 27 per dyad at 3000 ms each **Total exposure:** 81 sec/dyad	

Note: Table 3 reports the frequency and duration of A events (Experiment 2a) or B events (Experiment 2b). All conditions had five cue-outcome dyads. The baseline frequencies of A, B, and C events were 9 per dyad (45 total), and the baseline frequency for D events was 9. The baseline duration of all event types was 3000 ms. Each cell in Table 3 reports the frequency and duration of the event type (A or B) that was different from Baseline. All frequencies are per dyad. There was no baseline condition that actually received all the baseline values (i.e., 9 exposures/dyad to A, B, and C events for 3000 ms). Thus, Experiments 2a and 2b each used a 2 (Exposure level: high vs. low) x 3 (Trial variation: frequency vs. duration vs. adjusted frequency) factorial repeated-measures design.

Table 4. ***Design of Experiment 3***

		Duration (ms)		
		140	**560**	**2240**
Frequency	**3**		**Few**: 3 per dyad at 560 ms each **Total exposure:** 16.8 seconds	**Few Adjusted:** 3 per dyad at 2240 ms each **Total exposure:** 67.2 seconds
	12	**Short:** 12 per dyad at 140 ms each **Total exposure:** 16.8 seconds	**Baseline:** 12 per dyad at 560 ms each **Total exposure:** 67.2 seconds	**Long:** 12 per dyad at 2240 ms each **Total exposure:** 268.8 seconds
	48	**Many adjusted:** 48 per dyad at 140 ms each **Total exposure:** 67.2 seconds	**Many:** 48 per dyad at 560 ms each **Total exposure:** 268.8 seconds	

Note: Each of the seven conditions in Experiment 3 used ten dyads. Across conditions, the duration, frequency, or adjusted frequency of A events in each of the ten dyads was manipulated relative to a common Baseline condition. After exposure to the conditions described above, participants received a cued recall test on each of the ten dyads.

Table 5. ***Design of Experiment 4***

		Duration (ms)		
		333	**1000**	**3000**
Frequency	**3**		**Few**: 3 per dyad at 1000 ms each **Total exposure:** 3.0 seconds	**Few Adjusted:** 3 per dyad at 3000 ms each **Total exposure:** 9.0 seconds
	9	**Short:** 9 per dyad at 333 ms each **Total exposure:** 3.0 seconds		**Long:** 9 per dyad at 3000 ms each **Total exposure:** 27.0 seconds
	27	**Many adjusted:** 27 per dyad at 333 ms each **Total exposure:** 9.0 seconds	**Many:** 27 per dyad at 1000 ms each **Total exposure:** 27.0 seconds	

Note: The 12 conditions in Experiment 4 each contained five dyads. Across conditions, the duration, frequency, or adjusted frequency of A events in each of the conditions was manipulated relative to a common baseline frequency of 9 or a common baseline duration of 1000 ms. The baseline condition was not included in the experiment. After exposure to the six training conditions described above, participants received a cued recall test on each of the five dyads and, in a separate condition, a contingency rating test. In addition to the events described in the table, all conditions included nine B and C cell events per dyad and nine D events total. These additional events were all 1000 ms in duration. Thus, Experiment 4 used two 2 (Exposure level: high vs. low) x 3 (Trial variation: frequency vs. duration vs. adjusted frequency) factorial repeated-measures designs, one for cued recall testing and the other for contingency rating testing.

Table 5. ***Design of Experiment 5***

		Duration (ms)		
		110	**550**	**2750**
Frequency	**12**		**Few**: 12 at 550 ms each **Total exposure:** 6.6 seconds	**Few Adjusted:** 12 at 2750 ms each **Total exposure:** 33 seconds
	60	**Short:** 60 at 110 ms each **Total exposure:** 6.6 seconds	**Baseline:** 60 at 550 ms each **Total exposure:** 33 seconds	**Long:** 60 at 2750 ms each **Total exposure:** 165 seconds
	300	**Many adjusted:** 300 at 110 ms each **Total exposure:** 33 seconds	**Many:** 300 at 550 ms each **Total exposure:** 165 seconds	

Note: Each of the eight conditions in Experiment 5 contained ten dyads. Across conditions, the duration, frequency, or adjusted frequency of D events was manipulated relative to a common baseline condition. The values above describe the amount of exposure to D events. After exposure to the conditions described above, participants received either a cued recall test on each of the ten dyads or a test of relatedness for one randomly selected dyad. In addition to the D events described above, all conditions included 6 A events per dyad and 3 B events per dyad. All A and B events were 700 ms in duration. Thus, Experiment 5 used a 2 (Test: cued recall vs. relatedness rating) x 2 (Exposure level: high vs. low) x 3 (Trial variation: frequency vs. duration vs. adjusted frequency) + 1 (baseline) repeated measures design.

Figure Captions

Figure 1. Screen layout of cues and outcomes for each event. Only the left (right) pair of frames was presented on odd (even) numbered trials. A cue was present on A and B trials; an outcome was present on A and C trials.

Figure 2. Each panel reports the relationships between Frequency (top two rows) or Duration (bottom two rows) and relatedness ratings observed in Experiment 1. Each point illustrates a condition mean. The data from the Baseline conditions is repeated in all panels that used the same number of dyads (e.g., given three dyads, Cell A – Frequency, Cell B – Frequency, Cell A -Duration, and Cell B – Duration all reflect the same baseline data). Values in the top-left of each panel report the coefficient describing the linear relationship between frequency or duration, coded as described in the methods section, and contingency ratings. Vertical bars are bootstrapped 95% confidence intervals. Bootstrapping considered only observed scores within a condition. Hence, estimates ignored within-participant variability across conditions.

Figure 3. Condition mean relatedness ratings (± bootstrapped 95% confidence intervals) in Experiment 2a. Black bars show means for High Frequency, High Duration, and High Adjusted Frequency of A events. White bars report means for Low Frequency, Low Duration, and Low Adjusted Frequency of A events.

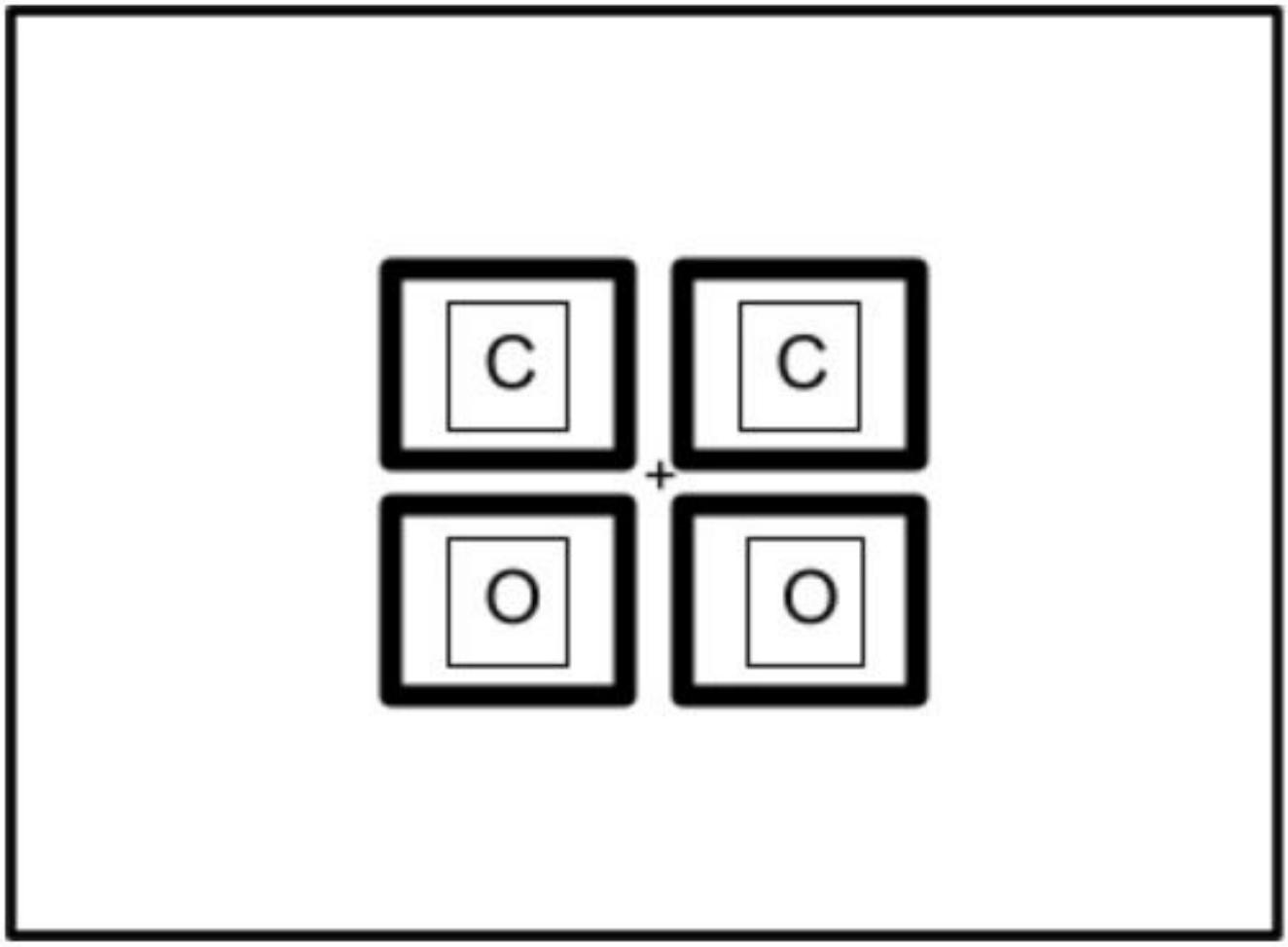
Figure 4. Condition mean relatedness ratings (± bootstrapped 95% confidence intervals) in Experiment 2b. Black bars show means for High Frequency, High Duration, and High Adjusted Frequency of B events. White bars report means for Low Frequency, Low Duration, and Low Adjusted Frequency of B events.

Figure 5. The left panel reports the relationships between Frequency (squares) or Adjusted Frequency (triangles) and mean memory scores (± 95% bootstrapped confidence interval) in Experiment 3. The right panel reports the relationship between Duration (stars) and mean memory scores (± 95% confidence interval). Each point illustrates a condition mean. The Baseline condition is repeated for all three

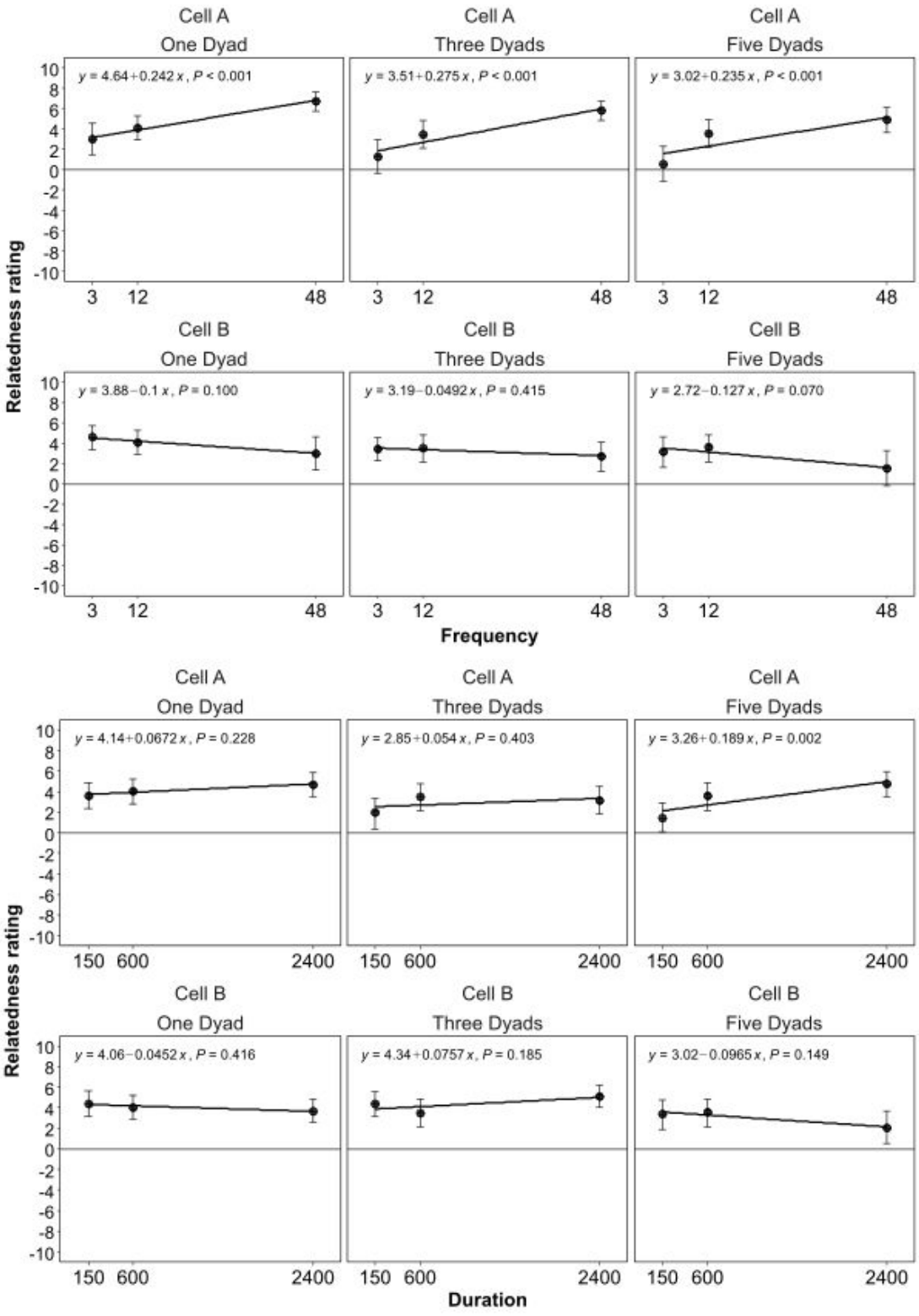
groupings of conditions using the same mean correct value. Values in the top-left of each panel report the estimated regression equation describing the linear relationship between frequency, adjusted frequency, or duration, coded as described in the methods section, and the number of items correctly recalled. The baseline condition is depicted once in each panel, but the mean was used to estimate the parameters of all three regression equations. All participants' scores contributed to the estimates shown in Figure 5, even if some of their data were excluded based on the elimination criteria.

Figure 6. Condition mean relatedness ratings (left panel) and memory scores (right panel) ± bootstrapped 95% confidence intervals in Experiment 4. Filled bars represent High exposure conditions, and unfilled bars represent Low exposure conditions. All participants' scores contributed to the estimates shown in Figure 6, even if some of their data were excluded based on the elimination criteria.

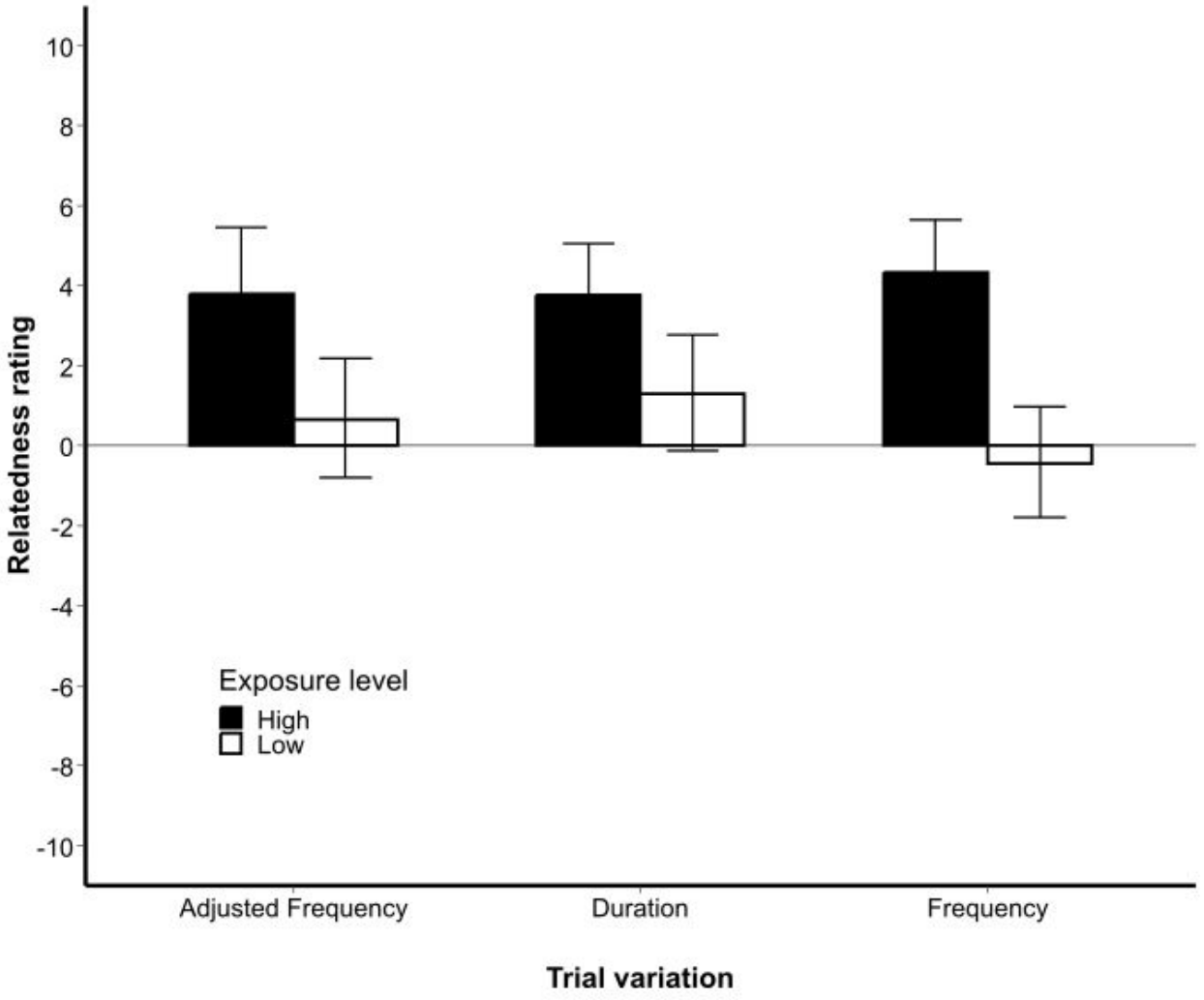
Figure 7. The top row reports the mean relatedness ratings, and the bottom row reports the mean recall scores observed in Experiment 5. The left panels report the results of frequency and adjusted frequency manipulations, and the right panels report the results of duration manipulations. All participants' scores are included in Figure 7, even if some of their data were excluded based on the elimination criteria.



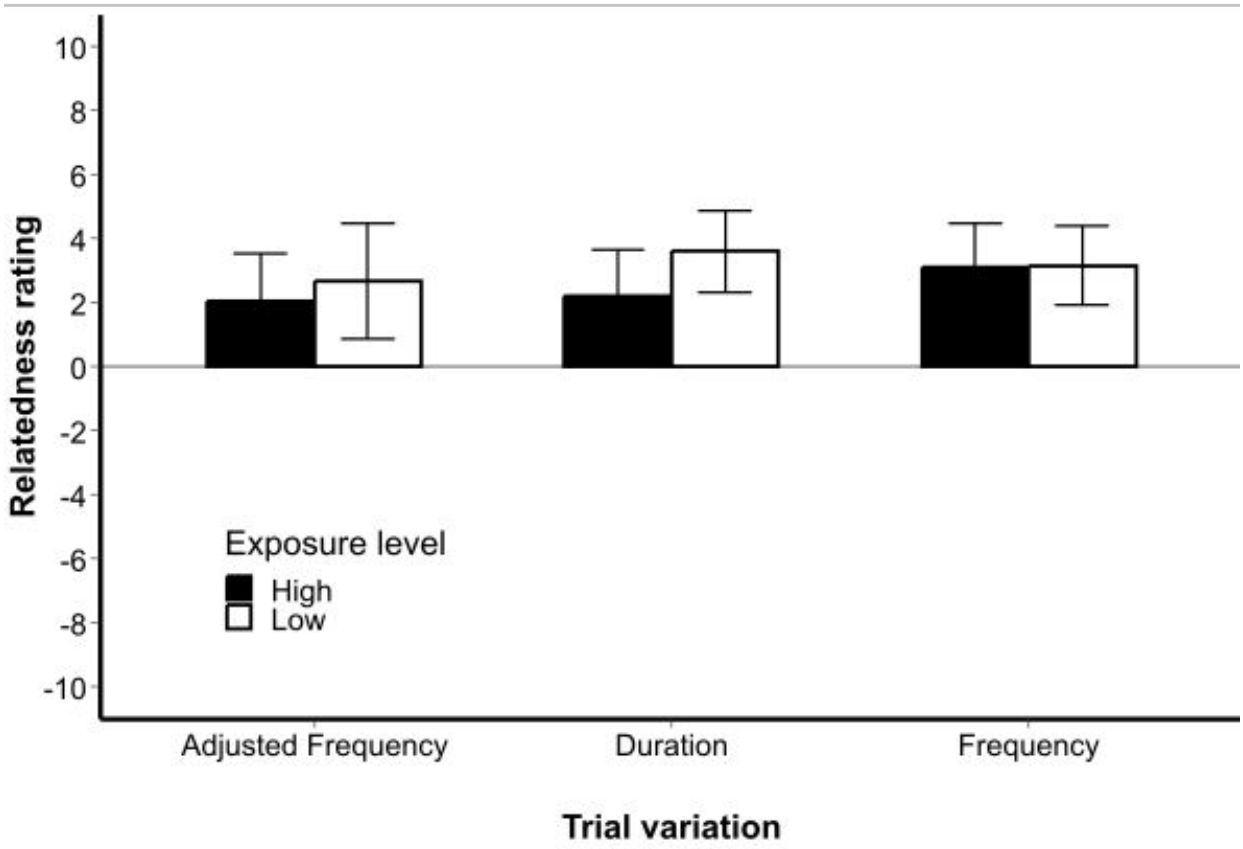
Witnauer et al., Figure 1



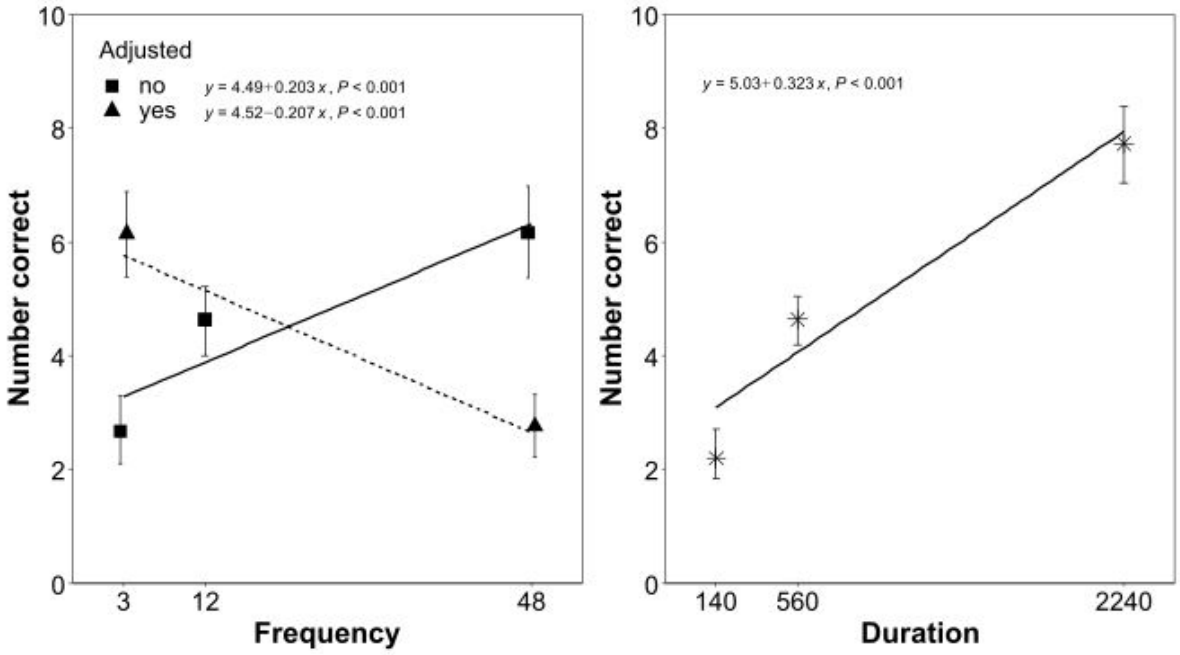
Witnauer et al., Figure 2



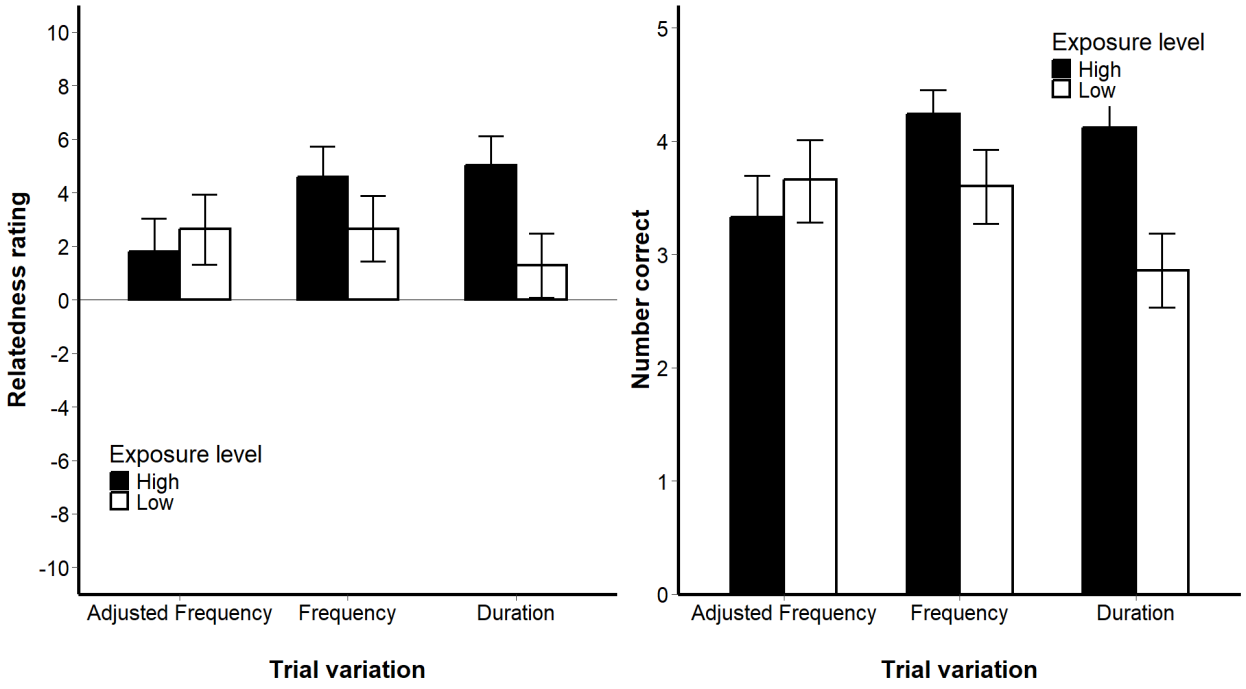
Witnauer et al., Figure 3



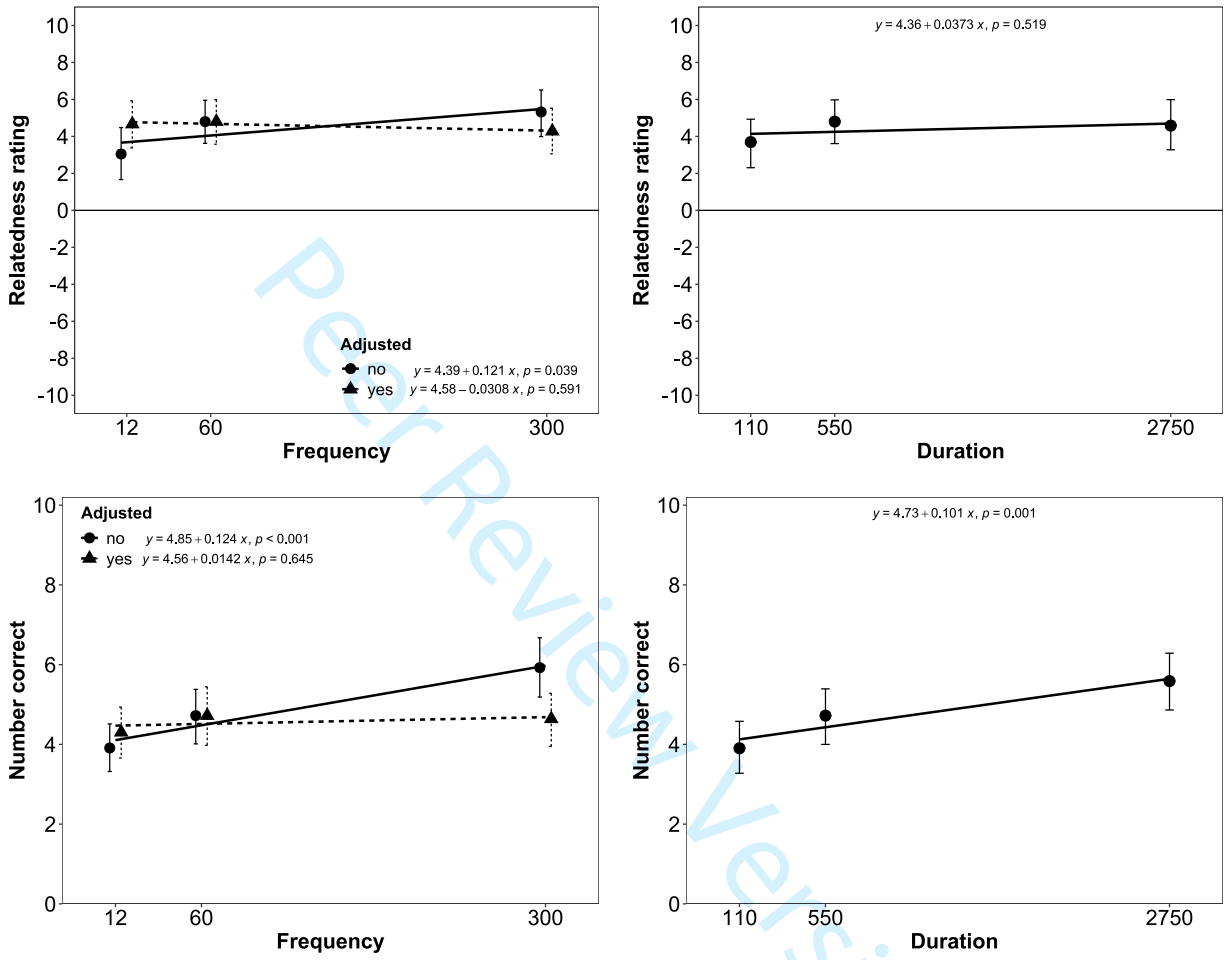
Witnauer et al., Figure 4



Witnauer et al., Figure 5



Witnauer et al., Figure 6



Witnauer et al., Figure 7

Appendix A

Participants were first instructed about the duration of the task and to turn off their cell phones.

> *Please turn off your cell phone, give your undivided attention to the screen, and refrain from changing programs while the experiment is in progress. The experiment lasts about 50 minutes, so please use the restroom before starting.*

After clicking "Continue," participants were instructed on how to take any needed breaks during the task.

> *This experiment requires your undivided attention. To receive compensation, you must: 1) Not change screens on your computer or look at your cellphone during the experiment. 2) If you have to take a break during the experiment, it must be immediately after answering a question and before you start the next set of stimuli. 3) You may not take more than two breaks. 4) No break can be more than 10 minutes, and break times do not count toward the 45-55 minutes this experiment actually runs. You must submit the HIT to MTurk within 120 minutes of accepting it.*

After clicking "Continue" again, participants were informed about the procedure.

> *In this experiment, you will be watching numerous series of rapidly presented shapes and drawings. After each series, a question screen will appear and you will be asked to rate the degree of relatedness between one shape and one drawing on a scale from -10 to +10. You will not know which shape and which drawing you will be asked about until the presentations are complete. As there will be many shapes and drawings, please try hard not to confuse them. Please keep your eyes on the cross in the center of the screen.*

The next screen presented information about the rating scale and was divided into chunks, with each chunk being presented for 10 seconds without the possibility of navigating away from the text and remaining in the study.

Chunk 1

A STRONG POSITIVE RATING should be given when the shape is always accompanied by the picture immediately below it, and when the absence of the shape is always accompanied by the absence of the picture immediately below it.

Chunk 2

A STRONG NEGATIVE RATING should be given when the shape is always presented without being accompanied by the picture below it, or when the picture is always presented without the shape having appeared immediately above it.

Chunk 3

A RATING OF ZERO should be given when the shape is equally presented with and without being accompanied by the picture below it.

Appendix B

Upon clicking a button to indicate consent, each participant received instructions on three separate screens, each of which was presented for 60 seconds.

Screen #1:

Please turn off your cell phone, give your undivided attention to the screen, and refrain from changing programs while the experiment is in progress. The experiment lasts about 27 minutes, so please use the restroom before starting.

Screen #2:

In this experiment, you will be watching numerous series of rapidly presented word pairs, the top in a foreign language and the bottom being its translation in English. After each series, a question screen will appear, and you will be asked to translate the foreign words into English. Responses delayed more than 10 seconds will result in your responses NOT being counted. Please keep your eyes on the cross in the center of the screen during the experiment.

Screen #3:

Please press continue when you are ready to begin the experiment.