Effects of Concurrent Load on Feature- and Rule-based Generalization in Human Contingency Learning

Andy J. Wills[1], Steven Graham[2], Zhisheng Koh[2], Ian P. L. McLaren[1], Matthew D. Rolland[1]

[1] University of Exeter

[2] National University of Singapore

Author Note

Andy J. Wills, Matthew D. Rolland, Ian P. L. McLaren, School of Psychology, University of Exeter, United Kingdom; Steven Graham, Zhisheng Koh, Department of Psychology, National University of Singapore

This research was supported by British Council grant RC105. The authors thank Jan De Houwer, Stephen Monsell, Nestor Schmajuk, David Shanks, Tom Verguts and an anonymous reviewer for their helpful comments, and Thomas Barrasin, Chris Longmore, Emily Mills and Charlotte Reilly for their assistance.

Correspondence concerning this article should be addressed to Andy J. Wills, School of Psychology, University of Exeter, Perry Road, Exeter. EX4 4QG. United Kingdom. E-mail: a.j.wills@ex.ac.uk

Revised revised manuscript for submission to JEP: ABP

5-Jan-11

Abstract

The effect of concurrent load on generalization performance in human contingency learning was examined in two experiments that employed the combined positive and negative patterning procedure of Shanks and Darby (1998). In Experiment 1 we tested 32 undergraduates and found that participants who were trained and tested under full attention showed generalization consistent with the application of an opposites rule (i.e., single cues signal the opposite outcome to their compound), whilst participants trained and tested under a concurrent cognitive load showed generalization consistent with surface similarity. In Experiment 2, we replicated the effect with 148 undergraduates, and provided evidence that it was the presence of concurrent load during training, rather than during testing, that was critical. Implications for associative, inferential, and dual-process accounts of human learning are discussed.

Keywords: rules; associative learning; generalization; working memory; deliberative processing.

Effect of Concurrent Load on Feature- and Rule-based Generalization in Human Contingency Learning

The distinction between deliberative and non-deliberative processing, under a variety of different names, is fundamental to the study of cognition. For example, theorists seek to distinguish between propositional and associative learning (McLaren, Green & Mackintosh, 1994; Mitchell, De Houwer & Lovibond, 2009), between analytic and nonanalytic categorization (Brooks, 1978; see also Ashby, Alfonso-Reese, Turken & Waldron, 1998), between automatic and intentional retrieval from memory (Jacoby, 1991; Mandler, 1980), and between intuitive and deliberate reasoning (Kahneman, 2003; see also Sloman, 1996). Deliberative processing is generally considered to go beyond surface similarity to extract causal (De Houwer & Beckers, 2003) or abstract (Shanks & Darby, 1998) structure, or to detect and correct irrational non-deliberative inferences (Kahneman, 2003). Deliberative thought processes are also often considered to be those that involve a degree of recurrence – in the sense that one goes through a series of intermediate stages to arrive at the final response (Milton & Wills, 2004). Another, related, way of capturing this idea of recurrence is to say that deliberative thought approximates the operation of a physical symbol system (Newell, 1980) – the ideas are related because certain recurrent, neural-like, structures have been shown to be able to implement a Universal Turing Machine (i.e. a general-purpose computer, Siegelmann & Sontag, 1995). Some argue that the ability to approximate a physical symbol system is what underlies the apparent discontinuity between human and non-human minds (Penn, Holyoak & Povinelli, 2008). Non-deliberative processing, on the other hand, is considered to be less effortful, less reliant on working memory resources, and to generalize on the basis of surface similarity (Kemler Nelson, 1984; J.D. Smith & Shapiro, 1989; E. E. Smith, Patalano & Jonides, 1998).

In the current studies we investigated the relationship between concurrent cognitive load and the extent to which people engage in deliberative processing when acquiring new information. It is known that people with comparatively large working memories learn some tasks more quickly than people with comparatively small working memories (Kyllonen & Stephens, 1990), and that concurrent cognitive load can disrupt the learning of, for example, within-compound associations (Aitken et al., 2001). In the current studies, we were interested primarily in the relationship between concurrent cognitive load and the *nature* of what is learned.

One related study is that by De Houwer and Beckers (2003). In their forward cue competition experiment, participants first observed (in a computer game scenario) that firing a particular weapon (A) was followed by the destruction of a tank. Later on, weapon A was fired simultaneously with a new weapon, B. This compound firing also led to the destruction of the tank. Participants were then asked about the causal status of weapon B with respect to the destruction of the tank. De Houwer and Beckers found that the imposition of a concurrent working memory load led to higher ratings for the extent to which B was considered to cause the tank's destruction, compared to a situation where the same contingencies were observed in the absence of a concurrent load. Their interpretation of these results was that the imposition of a concurrent load interfered with the deliberative processes required to work out that the causal status of B was uncertain (B is related to the outcome by contiguity, but B always occurs in the presence of A, and hence B's causal relation to the outcome is uncertain). Incidentally, De Houwer and Beckers (2003) found a significant effect of concurrent load only when load was applied during both the initial training phase and the subsequent rating phase. This is consistent with Aitken et al.'s (2001) earlier work, which employed concurrent load during training only, and found no significant effect of concurrent load on forward cue competition under these conditions.

In another related study, Waldron and Ashby (2001) demonstrated that concurrent working memory load retarded the acquisition of category structures definable in terms of a simple (single-attribute) rule, whilst the acquisition of category structures for which the rule was complex and non-intuitive was not significantly affected by concurrent load. Waldron and Ashby concluded that concurrent load interfered with the deliberative (rule-based) categorization process that would normally dominate in the simple-rule case, but that concurrent load left unaffected the non-deliberative processes underlying the acquisition of the more complex category structure (see also Filoteo, Lauritzen & Maddox, 2010).

One of the inherent difficulties in attempting to demonstrate a relationship between the availability of working memory resources and the extent to which learning proceeds deliberatively is to find a type of learning behavior that is unambiguously outside the scope of non-deliberative theories of cognition. For example, Nosofsky and Kruschke (2002) have argued that the results of Waldron and Ashby (2001) can be accounted for by a non-recurrent, non-deliberative, exemplar model (ALCOVE; Kruschke, 1992). The basis of Nosofsky and Kruschke’s argument is that concurrent load may be hypothesized to disrupt learned selective attention. This disruption will affect learning tasks in which selective attention is helpful (such as categories defined by a single attribute). Forward cue competition is also, at least in part, the result of learned selective attention (Kruschke, Kappeman & Hetrick, 2005; Wills, Lavric, Croft & Hodgson, 2007), so the idea that concurrent load disrupts selective attention might also, in principle, account for the forward cue competition results of De Houwer and Beckers (discussed above).

In the current studies, we examined the role of concurrent load in the learning and generalization task introduced by Shanks and Darby (1998). Our choice of task was motivated by the observation that the performance of a subset of participants in this task was, across all published discussions of the result that we were aware of, considered to be evidence for the role of deliberative processing in learning. This view had been expressed both by those who argue for a central role of associative processes in human learning (e.g. Cobos, Almaraz & Garcia-Madruga, 2003; Le Pelley, Oakeshott, Wills & McLaren, 2005), and those who argue that human learning is largely the result of conscious deliberative processes of inference (e.g. De Houwer & Beckers, 2002; Mitchell et al., 2009). Verguts and Fias (2009) have argued that the behavior of this subset of participants can be accounted for by a recurrent connectionist model – this argument is consistent with the position, outlined above, that recurrence is a characteristic feature of deliberative processing and that recurrent network architectures can implement general-purpose computational systems. Of course, the absence of published non-deliberative accounts of the Shanks and Darby (1998) results does not necessarily imply that plausible non-deliberative accounts are impossible – simply that none have thus far been publicly proposed. We return to this issue in the General Discussion, where it can be considered in relation to the data presented in the current paper.

The design and key result of Shanks and Darby (1998) is shown in Figure 1. Participants were asked to take the role of an allergist, attempting to predict which foods will cause an allergic reaction in a hypothetical patient, Mr. X. In Figure 1, letters stand for foods, "+" indicates the presence of an allergic reaction, and "-" indicates the absence of an allergic reaction. The training phase contained two complete negative patterning problems (e.g. A+, B+, AB-) and two complete positive patterning problems (e.g. C-, D-, CD+). Training also contained four incomplete patterning problems (e.g. participants see I+ and J+ but not IJ). The critical results concern participants' generalization to novel items, such as IJ, in the absence of feedback. For example, say you have observed that Mr. X develops an allergic reaction when he eats ice cream (I) and when he eats jelly (J). Do you predict the presence or absence of an allergic reaction when eating ice cream and jelly together?

A non-deliberative, surface similarity, process is likely to predict allergic reaction to IJ, as IJ is similar to both I and J, both of which produced an allergic reaction. A deliberative process, however, might detect that an *opposites* rule succinctly captures the information available during training – single foods produce the opposite reaction to their compounds. On this basis, IJ is predicted to not result in an allergic reaction, because this is the opposite outcome to that for I occurring on its own (and for J occurring on its own). As shown in Figure 1, Shanks and Darby found that participants who achieved a high level of accuracy during training showed generalization consistent with the application of an opposites rule, while participants who performed less well in training showed generalization consistent with surface similarity. Shanks and Darby's hypothesis (and ours) is that this transition in generalization reflects a transition from non-deliberative to deliberative processing during the course of training. We further hypothesize that, if opposites-rule and surface-feature generalization are indeed the products of deliberative and non-deliberative processing respectively, then the availability of working memory resources should determine whether opposites-rule or surface-feature generalization is seen (under the assumption that deliberative processing makes greater demands on working memory than non-deliberative processing).

Existing evidence could be employed, in a fairly indirect manner, to argue either for, or against, our hypothesis. On the one hand, Winman, Wennerholm, Juslin and Shanks (2005) demonstrated that opposites-rule generalization was related to performance on Raven's Progressive Matrices (RPM). RPM are considered to be a measure of general intelligence (g) and g appears to be related to working memory capacity (Conway, Kane & Engle, 2003). Hence one might argue that working memory capacity is likely to be positively correlated with opposites-rule generalization in the Shanks-Darby task and therefore that concurrent load, through limiting the availability of working memory resources, should lead to feature-based generalization. On the other hand, De Houwer and Vandorpe (2009) demonstrated performance consistent with opposites-rule generalization in the Implicit Association Test (IAT; Greenwald, McGhee & Schwartz, 1998). Although a matter of some debate (Fazio & Olson, 2003), the IAT (as its name suggests) is often considered to index non-deliberative processing. Similarly, generalization in tasks such as artificial grammar learning, which is often considered to be characterized by non-deliberative processing, can be quite abstract in nature (e.g. Beesley, Wills & Le Pelley, 2010). Also, opposites-rule generalization is related to participants demonstrating an inverse base-rate effect (Winman et al., 2005), yet demonstration of an inverse base-rate effect has been reported to be unaffected by a concurrent load (Lamberts & Kent, 2007).

In Experiment 1, we compared people learning and being tested on the Shanks and Darby task under full attention with those learning and being tested under a concurrent working memory load. We hypothesized that those trained and tested under concurrent load would show generalization more consistent with surface similarity, whilst those trained and tested in the absence of such a load would show generalization more consistent with the application of an opposites rule. Concurrent load was presented during test as well as during training in order to minimize the opportunity for concurrent load participants to extract the rule during the test phase – previous work on the forward blocking procedure (De Houwer & Beckers, 2003) indicated that this might be a possibility.

Experiment 1

Method

Participants, apparatus and stimuli. Thirty-two Exeter University students took part on a voluntary basis. They were tested individually in a quiet testing room using a PC connected to a 15” CRT monitor and to headphones, and running E-prime (Version 1.1). For half the participants the foods A-P (see Figure 1A) were, respectively, cheese, garlic, milk, mushrooms, sea food, red meat, olive oil, coffee, banana, eggs, orange squash, bread, avocado, peanuts, pasta, and chocolate. For the remaining participants, the foods assigned to A and B were swapped with those assigned to C and D, and similarly for E/F and G/H, for I/J and K/L, and for M/N and O/P.

Procedure. Participants were asked to assume the role of an allergist, predicting whether a hypothetical patient, Mr. X, would or would not develop an allergic reaction after eating a meal containing certain foods. Participants received up to eight blocks of training, each block comprising two presentations of the 18 training trial types shown in Figure 1A, in a random order. At the end of each block, allergy prediction performance in that block was assessed against a criterion of 32 correct responses (out of 36, 89% accuracy). Participants meeting or exceeding this criterion proceeded immediately to the test phase; those that did not continued with training. The transition between training blocks was not signaled to participants, but the transition to the test phase was signaled (by the presentation of instructions indicating feedback would no longer be provided). The test phase comprised two presentations of the 24 test trial types shown in Figure 1A, in a random order.

Participants were randomly assigned to either the *load* or the *no-load* condition of the experiment. In the load condition, each trial (in both training and in test) began with the presentation of six, different, single-digit, randomly selected, numbers over headphones at 330ms intervals. After the presentation of the 6th digit, a fixation cross appeared on the screen for 500ms, followed by a blank screen for 500ms. Food names were then presented on the screen, and participants pressed a key to indicate whether or not Mr. X would suffer an allergic reaction. No time limit was set for these responses. During the training phase, each trial was followed by a feedback message of 1500ms duration (e.g. “Correct! Mr. X developed an allergic reaction”). No feedback was given during the test phase. In both the training phase and the test phase, each trial ended with participants being presented with a single digit on the screen (the probe), to which they were expected to respond by reporting the digit that immediately followed the probe in the most recently presented set of six digits. For example, if the participant had just heard, "6, 8, 2, 7, 9, 3", and the probe digit was 8, then the correct answer was "2". Participants reported their decision by pressing the corresponding number key on the computer keyboard. No feedback was given on the number responses, and no time limit was imposed. The next trial began 500 ms after the participant pressed a number key.

The procedure for the no-load condition differed from the procedure for the load condition in that the auditory presentation of digits was removed and replaced by a blank interval of the same duration. A digit was still visually presented at the end of each trial, but the task in the no-load condition was to press the key on the keyboard corresponding to the visually presented digit (e.g. to press 9 if the presented digit was 9). The no-load condition was implemented in this way in order to approximately match the timing and motor response requirements of the load condition. Of course, as intended, the two conditions were not matched on the difficulty of the secondary task.

Practice trials. In both the load and the no-load condition, the experiment proper was preceded by 10 practice trials on the digit task. In the load condition, these practice trials were identical to the test phase trials of the load condition, except that the meal presentation was replaced by an instruction to press one of the two response keys. In the practice task for the no-load condition, the auditory presentation of digits was replaced by a blank interval of the same duration, and the practice task involved pressing a key on the keyboard corresponding to the visually presented digit.

Results and Discussion

Seven participants from the load condition, and three participants from the no-load condition, failed to meet the training criterion within the eight blocks available and were excluded from further analysis. After these exclusions, proportion correct in the final training block was similar for the concurrent load participants (mean = .94, SD = .04), and the no-load participants (mean = .93, SD = .04), $t(20) < 1$. Unsurprisingly, the load participants took longer to reach criterion (mean = 6.9 blocks , SD = 1.1) than the no-load participants (mean = 4.2, SD = 1.5), $t(20) = 4.80$, $p < .0005$. Accuracy on familiar test items was investigated via ANOVA, with a between-subject factor of concurrent load (present vs. absent), and a within-subject factor of test item type (positive items vs. negative items; where positive and negative indicate the presence and absence of an allergic reaction to the food during training). Although there was a main effect of test item type, $F(1,20) = 873.8$, $p < .0005$, neither the main effect of concurrent load, $F(1,20) < 1$, nor the interaction between concurrent load and test item type, $F(1, 20) = 2.64$, were significant. Hence, no statistically reliable effects of concurrent load were found on test items familiar to participants. This is consistent with the hypothesis that responses to familiar test items in this study are the product of a relatively automatic memory retrieval process. The mean proportion of allergy responses to positive familiar test items was .93, whilst the mean proportion of allergy responses to negative familiar test items was .08.

As Figure 2A illustrates, participants in the no-load condition responded to the novel compounds (MN, IJ) in a manner consistent with the application of an opposites rule, whilst participants in the load condition responded in a manner consistent with surface-similarity generalization. ANOVA confirms the presence of a significant interaction between concurrent load (present vs. absent) and stimulus type (MN vs. IJ), $F(1,20) = 27.8$, $p < .0005$; the main effect of concurrent load, $F(1, 20) = 1.12$, and the main effect of stimulus type, $F(1, 20) = 2.34$, were not significant. Similarly, and as illustrated in Figure 2B, participants in the no-load condition responded in a manner consistent with the application of an opposites rule to K/L and O/P, whilst participants in the load condition responded in a manner consistent with surface-similarity generalization. ANOVA confirmed the presence of a significant interaction between concurrent load (present vs. absent) and stimulus type (K/L vs. O/P), $F(1,20) = 18.2$, $p < .0005$; the main effect of concurrent load, $F(1, 20) < 1$, and the main effect of stimulus type, $F(1,20) = .69$, were not significant.

Further investigation of the interaction illustrated in Figure 2A revealed that the proportion of allergy responses was significantly higher for MN than for IJ in the absence of concurrent load, $t(12) = 7.58$, $p < .005$. In the presence of concurrent load, IJ is higher than MN, but this difference is not significant, $t(8) = 1.84$, $p = .10$. Further investigation of the interaction illustrated in Figure 2B revealed that the proportion of allergy responses was significantly higher for K/L than for O/P in the absence of concurrent load, $t(12) = 4.56$, $p < .005$. In the presence of concurrent load, K/L was lower than O/P but this difference was not significant, $t(8) = 1.74$, $p = 0.12$.

Finally, it is possible to combine the four probabilities IJ, MN, K/L and O/P into a single index of opposites-rule versus surface-similarity generalization. Computing (MN - IJ + mean[K,L] – mean[O,P]) / 2, results in an “rule following” index that ranges from -1 (perfect feature-based generalization) to + 1 (perfect opposites-rule generalization). In the absence of concurrent load during training, the mean rule-following index is .65 (SD = 0.33), which is significantly greater than zero, $t(12) = 7.20$, $p < .005$. In the presence of concurrent load during training, the mean rule-following index is -.33 (SD = .50), which is not significantly less than zero, $t(8) = 1.99$, $p = .08$.

In summary, concurrent load reverses the order of allergy prediction proportions for novel test items, compared to conditions of full attention (no concurrent load). In the absence of concurrent load, the order of allergy prediction proportions implies rule-based generalization, whilst in the presence of concurrent load, the order of allergy prediction proportions implies similarity-based generalization. This pattern in the means is supported by statistically significant interactions in the novel test items, and no reversal or statistically significant interaction in the familiar test items. However, the simple effects (IJ > MN; K/L < O/P) in the presence of concurrent load fall short of significance, possibly due to the relatively small sample size. Hence, one goal of Experiment 2 was to attempt to provide a larger-sample replication of the results of Experiment 1.

The other goal of Experiment 2 was to identify the locus of the effect of concurrent load. In the load condition of Experiment 1, concurrent load was applied during both training and test. In the no-load condition of Experiment 1, no concurrent load was applied in either phase. The difference between the load and no-load conditions of Experiment 1 might therefore be as a result of concurrent load during training, concurrent load during test, or both. In Experiment 2 we factorially manipulated the presence of concurrent load during the training phase, with the presence of concurrent load during the test phase.

Experiment 2

Method

Participants, apparatus and stimuli. One hundred and forty eight National University of Singapore students took part in return for course credit. They were tested in groups of between 10 and 20 students, with each participant seated at a different PC. All PCs were of an identical make and model, and each was connected to a 15” CRT monitor, and to headphones. The E-prime package (Version 1.1) was employed. The stimuli were identical to those employed in Experiment 1.

Procedure. Each participant was randomly allocated to one condition of the 2 x 2 factorial between-subjects design of this experiment. The two factors were concurrent load during the training phase (present or absent) and concurrent load during the test phase (present or absent). The crossing of these factors led to four experimental conditions – *No Load* (no concurrent load during training and no concurrent load during test), *Train Load* (concurrent load during training, but not during test), *Test Load* (concurrent load during test but not during training), and *Both Load* (concurrent load during training and concurrent load during test). The procedure for the Both Load condition was identical to the procedure in the Load condition of Experiment 1. The Train Load condition was identical to the Both Load condition, except that digits were not presented over headphones during the test phase (they were replaced by a silent pause of the same duration) and participants responded to the visually presented digits during the test phase by pressing the corresponding key (e.g. pressing “3” if the number 3 appeared). In a corresponding manner, the Test Load condition was identical to the Both Load condition, except that digits were not presented over headphones during the training phase, and participants responded to the visually presented digits during the training phase by pressing the corresponding key.

The procedure for the No Load condition was identical to the procedure for the No Load condition in Experiment 1, except for the ten trials of practice on the digit task at the beginning of the experiment. In Experiment 2, all conditions began with ten trials of practice on the digit task (in Experiment 1, the No Load condition had practice trials, but these simply involved pressing the key corresponding to the digit shown on screen, and no digits were presented over headphones).

Results and Discussion

Thirty seven participants failed to meet the training criterion within the eight blocks available, and were excluded from further analysis (leaving 34, 23, 25, and 29 participants in the No Load, Train Load, Test Load, and Both Load conditions, respectively). After these exclusions, proportion correct in the final training block was similar for those who encountered concurrent load during training (mean = .92, SD = .02), and those who did not (mean = .93, SD = .03), $t(109) = 1.89$, $p > .05$. Those encountering concurrent load during training took longer to reach criterion (mean = 6.0 blocks, SD = 1.3) than those who did not (mean = 4.8, SD = 1.8), $t(109) = 4.23$, $p < .0005$. Accuracy on familiar test items was investigated via ANOVA, with a between-subject factor of concurrent load during training (present vs. absent), another between-subject factor of concurrent load during test (present vs. absent), and a within-subject factor of test item type (positive items vs. negative items; where positive and negative indicate the presence and absence of an allergic reaction to the food during training). Although there was a main effect of test item type, $F(1,107) = 2539$, $p < .0005$, no other main effect or interaction approached significance, $F(1, 107) < 1.9$. Hence, no statistically reliable effects of concurrent load were found on test items familiar to participants. The mean proportion of allergy responses to positive familiar test items was .90, whilst the mean proportion of allergy responses to negative familiar test items was .11.

As in Experiment 1, the result of central interest is the generalization performance on the novel test items. Considering the novel compounds (IJ, MN) first, an ANOVA with two between-subject factors (load during training: present or absent; load during test: present or absent) and one within-subject factor (compound type: IJ or MN) revealed a significant interaction between compound type and concurrent load during training, $F(1, 107) = 12.93$, $p < .0005$. This interaction is illustrated in Figure 3A. No main effects, and no other interactions, approached significance, $F(1, 107) < 1.5$. In particular, there was no significant main effect of concurrent load during test, and no interaction of load during test with any other factor. Similarly, consideration of the novel elements during the test phase revealed a significant interaction between element type (K/L vs. O/P) and load during training, F(1, 107)

= 22.75, $p < .0005$. This interaction is illustrated in Figure 3B. Again, no main effects, and no other interactions, were significant, $F(1, 107) < 2.5$.

Further investigation of the interaction illustrated in Figure 3A revealed that the proportion of allergy responses was significantly higher for MN than for IJ in the absence of concurrent load during training, $t(58) = 3.26$, $p < .005$. In the presence of concurrent load during training, IJ was higher than MN, and this difference is significant as a one-tailed test, $t(51) = 1.72$, $p < .05$. A one-tailed test is appropriate here, given that the direction of the trend is predicted on the basis of both theory and the results of Experiment 1[1]. Further investigation of the interaction illustrated in Figure 3B reveals that the proportion of allergy responses was significantly higher for K/L than for O/P in the absence of load during training, $t(58) = 2.85$, $p < .01$, whilst K/L was significantly lower than O/P in the presence of load during training, $t(51) = 3.55$, $p < .005$.

Finally, as in Experiment 1, it is possible to combine the four probabilities IJ, MN, K/L and O/P into a single index of opposites-rule versus surface-similarity generalization. Computing (MN - IJ + mean[K,L] – mean[O,P]) / 2, results in an “rule following” index that ranges from -1 (perfect feature-based generalization) to + 1 (perfect opposites-rule generalization). In the absence of concurrent load during training, the mean rule-following index is .26, which is significantly greater than zero, $t(58) = 3.41$, $p < .005$. In the presence of concurrent load during training, the mean rule-following index is -0.22, which is significantly less than zero, $t(51) = 2.69$, $p < .05$.

In summary, it seems that concurrent load during the training phase, rather than concurrent load during the test phase, is the critical factor underlying whether generalization is consistent with an opposites rule or with surface similarity.

General Discussion

In Experiment 1, we trained all participants to a high criterion of accuracy (89%). Half of the participants were trained and tested under full attention, whilst the remaining participants were trained and tested under a concurrent cognitive load. Those not under concurrent load showed generalization more consistent with an abstract rule, whilst those under concurrent load showed generalization more consistent with surface similarity. Experiment 2 replicated this effect, and demonstrated that it was load during training, rather than load during test, that was critical in determining the form of generalization observed.

The learning task employed in these studies was introduced by Shanks and Darby (1998). The task is unusual in that there is, amongst the published discussions of this task we are aware of, broad agreement that the opposites-rule generalization seen in this task is evidence for deliberative processes in human contingency learning (e.g. those thought processes outside the scope of non-recurrent associative models; Cobos et al., 2003; De Houwer & Beckers, 2002; Le Pelley et al. 2005; Mitchell et al., 2009; Verguts & Fias, 2009). The current level of published agreement concerning the Shanks and Darby task contrasts with the contentious nature of some other forms of putatively deliberative behavior that have been reported in adults (e.g. DeCaro et al., 2008 vs. Tharp & Pickering, 2009), in infants (e.g. Marcus, Vijayan, Rao & Vishton, 1999 vs. McClelland & Plaut, 1999), and in rats (Beckers, Miller, De Houwer & Urushiara, 2006 vs. Haselgrove, 2010).

Of course, a consensus amongst published discussions that opposites-rule generalization is deliberative does not necessarily imply that an account of “opposites-rule” generalization is impossible within an associative / connectionist architecture. In the Introduction, we mentioned the connectionist model of the Shanks-Darby results presented by Verguts & Fias (2009) and argued that their model is consistent with a deliberative account in the sense that it makes use of a recurrent network architecture. Deliberative processes are commonly considered to involve a degree of recurrence (e.g. Milton & Wills, 2004), and certain recurrent network architectures are known to have the computational power of Turing machines (Siegelmann & Sontag, 1995). It is therefore difficult to conceive of any result that would lead to the rejection of recurrent network architectures as a class of explanation for human behavior.

A perhaps more substantive issue is that one might be able to explain opposites-rule generalization within relatively simple associative architectures. The theoretical possibility of such an account is highlighted by the recent publication of relatively simple associative models of a range of effects in learning previously claimed to be outside the scope of such models (e.g. Haselgrove, 2010; Schmajuk, 2010). In the context of the current data, it might be possible, for example, to develop an account based on the concept of outcome-mediated generalization (Dickinson, personal communication). During training, A and B are both individually paired with the outcome, so presentation of the AB compound may lead to the activation of a super-normal outcome representation ("++"). That super-normal outcome representation would then itself become associated to an absence-of-outcome representation. The novel test phase compound IJ is predicted to be followed by no outcome because both I and J were associated with an outcome during training, and thus the IJ compound activates the super-normal outcome representation which in turn activates a representation of outcome absence. It seems that such an account must assume that activation of the super-normal outcome representation has less control over responding than the activation of the no-outcome representation by the super-normal outcome representation. This would seem to imply, in turn, some process that gives greater weight to super-normal (and presumably also sub-normal) outcome representations due to their high diagnosticity in the context of the training phase. Similar processes have been previously hypothesized at the level of cue representations (e.g. Mackintosh, 1975).

Although such an account is potentially promising, it needs further elaboration before it could encompass the results presented in the current article. Concurrent load during training results in the opposites-rule generalization observed to novel stimuli reversing to a pattern of feature-based generalization to those novel stimuli. At the same time, concurrent load during the test phase does not affect the pattern of generalization to novel test items. Both patterns of generalization (opposites rule and feature-based) were found in the context of near-perfect accuracy on the training items. Thus, any proposed mechanism faces a number of challenges. It should be unaffected by load at test. It should be able to produce high accuracy under load during training. It should be able to produce the correct pattern of generalization performance both with and without load during training – noting that the effect of load at training is to reverse the ordinal pattern of results at test. Finally, it needs to provide some explanation of why concurrent load would change the way the problem was solved.

The explanation we tentatively offer for our results is that participants initially approach the training phase through a process of exemplar storage and retrieval. This is a relatively non-deliberative process (which is not to say it is necessarily entirely automatic, see e.g. Logan & Etherton, 1994). Under conditions of full attention during training, participants notice there is a non-intuitive rule that substantially reduces the number of things one has to remember. At limit, all one needs to remember is the rule that compounds predict the opposite to their elements, and that the compounds CD, EF, and KL make Mr. X sick. Everything else can be derived – for example, A on its own will make Mr.X sick, because A is not in any of the three compounds that make him sick (CD, EF, KL), and compounds predict the opposite to their elements.

This process of rule extraction is assumed to be relatively deliberative and effortful, requiring as it does the generation of verbal hypotheses and then testing them against subsequently presented training items. If the process of rule extraction is assumed to be relatively deliberative and effortful, then it is not unreasonable to assume it might be selectively disrupted by concurrent load, whilst concurrent load might be expected to leave the processes of exemplar storage and retrieval relatively intact. Generalization to novel test items will depend on the nature of the representations developed during training. Under concurrent load, the representations are exemplars, and generalization would therefore be expected to be on the basis of surface similarity. Under full attention, the opposites rule is extracted, and generalization might therefore be expected to be on the basis of that opposites rule. Under this account, the fact that concurrent load during test has no detectable effect suggests that whilst the extraction of an opposites rule is highly effortful, the application of an already-extracted rule is not particularly effortful. The explanation we offer for our results falls amongst the broad class of explanations that assume cognition in adult humans is the product of at least two systems – one system that is deliberative, and perhaps approximates the function of a physical symbol system (Newell, 1980), and another system that is non-deliberative, and which might be approximated by a simple associative system. Explanations of this general class include those forwarded by Ashby et al. (1998), Brooks (1978), McLaren et al. (1994, see also Spiegel & McLaren, 2006; Jones & McLaren, 2009; Livesey & McLaren, 2009) and Sloman (1996).

In addition to simple associative, and dual-process, accounts of human cognition, a third class of theory is that all learning is the product of a deliberative system (e.g. inferential accounts; Mitchell et al., 2009). In relation to the results of the current experiments, inferential accounts would presumably assume that not only opposites-rule performance, but also surface similarity performance, was the product of an inferential process in this task. In order to predict any effect of concurrent load, such an account must assume that opposites-rule performance is more effortful than surface similarity performance – perhaps because participants arrive with a pre-experimental hypothesis that similar meals lead to similar outcomes, whilst the non-intuitive opposites-rule hypothesis is only arrived at through a relatively effortful process of hypothesis-testing during training. Concurrent load presumably interferes with this process, leaving the participant with just memory for the examples (which is employed for familiar test items) and a pre-experimental surface similarity hypothesis (which is employed for novel test items). As in a dual-process account, in order to account for the apparent ineffectiveness of load during test, one needs to assume that the application of a hypothesis is less effortful than its extraction.

In summary, an inferential explanation seems to need to assume both the presence of a relatively non-deliberative exemplar storage and retrieval process, and that certain types of inferential process are also relatively non-deliberative (e.g. inferences on the basis of a pre-experimental hypothesis about novel test items). In other words, an inferential explanation seems to need to assume the presence of not only relatively non-effortful forms of learning and retrieval, but also the presence of relatively non-effortful forms of inference. When expressed in those terms, an inferential account seems to largely converge with the dual-process we offer above. This contention, however, requires further research – research that will probably need to involve the application of formal models to a broader range of data than is currently typical (for further discussion of the importance of such an approach, see Wills & Pothos, 2011).

The results of the current study seem to lead to a couple of predictions amenable to further test. First, De Houwer and Vandorpe (2009) demonstrated the presence of opposites-rule performance in the Implicit Association Task (IAT). Although the IAT is conducted under conditions of full attention, performance on the IAT is often considered to reflect non-deliberative processing. However, the current results raise the possibility that the application of concurrent load to De Houwer and Vandorpe's IAT procedure might lead to a reversal from opposites-rule to surface-feature generalization. If such a pattern of results were observed, it would further complicate the interpretation of the IAT as a measure of non-deliberative responding (cf. Fazio & Olson, 2003). A second prediction that can be made on the basis of the current results is that the failure of Lamberts and Kent (2007) to find an effect of concurrent load on the inverse base-rate effect might be due to their decision to apply concurrent load only during the test phase. By analogy to the pattern of results reported in the current paper, one might hypothesize that it is the acquisition of the knowledge during the training phase, rather than demonstration of that knowledge in the test phase, that is particularly effortful in the inverse base-rate task. Concurrent load during training might therefore be expected to be more effective in modulating performance in this task.

Another avenue for further research might be to investigate whether there are conditions under which non-human animals would demonstrate generalization performance consistent with an opposites rule in this task. A number of recent studies have examined the extent to which generalization might be qualitatively different in human and non-human animals, but the results have been mixed. For example, Couchman, Coutinho and Smith (2010) reported that apes showed a greater incidence of overall similarity generalization, and a lower incidence of generalization on the basis of a single dimension, than humans. However, Wills et al. (2009) found no differences between humans, squirrels and pigeons in the incidence of overall similarity generalization, and found that the incidence of overall similarity sorting in humans and pigeons was affected by stimulus factors (e.g. spatial separate vs. spatially contiguous features) in a similar way in both species. Part of the problem may be that, whilst generalization on the basis of overall similarity is often considered to be the product of non-deliberative processes in humans (e.g. Kemler Nelson, 1984), this conclusion is rendered more complex by the increasing number of demonstrations that overall similarity generalization in humans can also be the product of a deliberative process (Milton, Longmore & Wills, 2008; Milton, Wills & Hodgson, 2009; Wills, Longmore

& Milton, 2011). The relatively broad agreement that opposites-rule generalization in the Shanks-Darby task is evidence for deliberative processing in humans may make it particularly suitable as a device for examining the extent of deliberative processing in non-human animals.

In summary, we have reported two demonstrations that the availability of working memory resources appears to be an important determinant of how we generalize from known instances to novel items. In our studies, information acquired under full attention seems to lead to generalization more consistent with the application of an abstract rule, whilst information acquired in the presence of concurrent load seems to lead to generalization more consistent with surface similarity.

Footnote

[1] Alternatively one can make use of the information provided by Experiment 1 within a two-tailed test by employing simple meta-analysis. Meta-analysis of Experiments 1 and 2 using the two-tailed probabilities (see Rosenthal, 1978) confirms that IJ is significantly greater than MN overall, $Z_1 = 1.63$, $Z_2 = 1.69$, $Z_{combined} = 2.34$, $p = .02$.

References

Aitken, M. R. F., M. J. W. Larkin, et al. (2001). Re-examination of the role of within-compound associations in the retrospective revaluation of causal judgements. *Quarterly Journal of Experimental Psychology, 54B*, 27-52.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442-481.

Beckers, T., Miller, R. R., De Houwer, J., & Urushihara, K. (2006). Reasoning rats: Forward blocking in Pavlovian animal conditioning is sensitive to constraints of causal inference. *Journal of Experimental Psychology-General, 135*, 92-102.

Beesley, T., Wills, A. J., & Le Pelley, M. E. (2010). Syntactic transfer in artificial grammar learning. *Psychonomic Bulletin and Review, 17*, 122-128.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization*. Hillsdale, N.J.: Erlbaum.

Cobos, P. L., Almaraz, J., & Garcia-Madruga, J. A. (2003). An associative framework for probability judgment: An application to biases. *Journal of Experimental Psychology-Learning Memory and Cognition, 29*, 80-96.

Conway, A. R. A., & Engle, R. W. (1994). Working-Memory and Retrieval - a Resource-Dependent Inhibition Model. *Journal of Experimental Psychology-General, 123*, 354-373.

Couchman, J. J., Coutinho, M. V. C., & Smith, J. D. Rules and Resemblance: Their Changing Balance in the Category Learning of Humans (Homo sapiens) and Monkeys (Macaca mulatta). *Journal of Experimental Psychology-Animal Behavior Processes, 36*, 172-183.

DeCaro, M. S., Thomas, R. D., & Beilock, S. L. (2008). Individual differences in category learning: Somtimes less working memory capacity is better than more. *Cognition, 107*, 284-294.

De Houwer, J., & Beckers, T. (2002). A review of recent developments in research and theories on human contingency learning. *Quarterly Journal of Experimental Psychology Section B-Comparative and Physiological Psychology, 55*, 289-310.

De Houwer, J., & Beckers, T. (2003). Secondary task difficulty modulates forward blocking in human contingency learning. *Quarterly Journal of Experimental Psychology, 56B*, 345-357.

De Houwer, J., & Vandorpe, S. (2009). Using the Implicit Association Test as a Measure of Causal Learning Does not Eliminate Effects of Rule Learning. *Experimental Psychology, 57*, 61-67.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54*, 297-327.

Filoteo, J. V., Lauritzen, S., & Maddox, W. T. (2010). Removing the frontal lobes: The effects of engaging executve functions on perceptual category learning. *Psychological Science, 21*, 415-423.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464-1480.

Haselgrove, M. (2010). Reasoning rats or associative animals? A Common-element analysis of the effects of additive and subadditive pretraining on blocking, *Journal of Experimental Psychology: Animal Behavior Processes.* In press.

Jacoby, L. L. (1991). A Process Dissociation Framework - Separating Automatic from Intentional Uses of Memory. *Journal of Memory and Language, 30*, 513-541.

Jones, F.W. and McLaren, I.P.L. (2009). Human Sequence Learning Under Incidental and Intentional Conditions. *Journal of Experimental Psychology: Animal Behavior Processes, 35,* 538-553.

Kahneman, D. (2003). A perspective on judgment and choice - Mapping bounded rationality. *American Psychologist, 58*, 697-720.

Kemler Nelson, D. G. (1984). The effects of intention on what concepts are acquired. *Journal of Verbal Learning and Verbal Behavior, 23*, 734-759.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology-Learning Memory and Cognition, 31*, 830-845.

Kyllonen, P. C., & Stephens, D. L. (1990). Cognitive abilities as determinants of success in acquiring logic skill. *Learning and Individual Differences, 2*, 129-160.

Lamberts, K., & Kent, C. (2007). No evidence for rule-based processing in the inverse base-rate effect. *Memory & Cognition, 35*, 2097-2105.

Le Pelley, M. E., Oakeshott, S. M., Wills, A. J., & McLaren, I. P. L. (2005). The outcome specificity of learned predictiveness effects: Parallels between human causal learning and animal conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 31*, 226-236.

Livesey, E.J. and McLaren I. P.L. (2009). Discrimination and Generalization Along a Simple Dimension: Peak Shift and Rule-Governed Responding. *Journal of Experimental Psychology: Animal Behavior Processes, 35,* 554-565.

Logan, G. D. & Etherton, J. L. (1994). What is learned in automatization? The role of attention in constructing an instance. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20,* 1022-1050.

McClelland, J. L., & Plaut, D. C. (1999). Does generalization in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences, 3*, 166-168.

McLaren, I.P.L., Green, R.E.A. & Mackintosh, N.J. (1994). Animal learning and the implicit/explicit distinction. In N.C. Ellis (Ed.) *Implicit and explicit learning of languages*. Academic Press.

Mackintosh, N.J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82*,276-298.

Mandler, G. (1980). Recognizing - the Judgment of Previous Occurrence. *Psychological Review, 87,* 252-271.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science, 283*(5398), 77-80.

Milton, F., Longmore, C. A., & Wills, A. J. (2008). Processes of overall similarity sorting in free classification. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 676-692.

Milton, F., & Wills, A. J. (2004). The influence of stimulus properties on category construction. *Journal of Experimental Psychology: Learning, Memory & Cognition, 30*, 407-415.

Milton, F., Wills, A. J., & Hodgson, T. L. (2009). The neural basis of overall similarity and single-dimension sorting. *NeuroImage, 46*, 319-326.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32*, 183-+.

Newell, A. (1980). Physical Symbol Systems. *Cognitive Science, 4*, 135-183.

Nosofsky, R. M., & Kruschke, J. K. (2002). Single-system models and interference in category learning: Commentary on Waldron and Ashby (2001). *Psychonomic Bulletin & Review, 9*, 169-174.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences, 31*, 109-+.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85,* 185-193.

Schmajuk, N.A. (2010). *Mechanisms in Classical Conditioning.* Cambridge: Cambridge University Press.

Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes, 24*, 405-415.

Siegelmann, H. T., & Sontag, E. D. (1995). On the Computational Power of Neural Nets. *Journal of Computer and System Sciences, 50*, 132-150.

Sloman, S. A. (1996). The Empirical Case for Two Systems of Reasoning. *Psychological Bulletin, 119*, 3-22.

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition, 65*(2-3), 167-196.

Smith, J. D., & Shapiro, J. H. (1989). The occurrence of holisitc categorization. *Journal of Memory and Language*, *28,* 386-399.

Spiegel, R. and McLaren, I.P.L. (2006). Associative Sequence Learning in Humans. *Journal of Experimental Psychology: Animal Behavior Processes, 32*, 156-63.

Tharp, I. J., & Pickering, A. D. (2009). A note on DeCaro, Thomas, and Beilock (2008): Further data demonstrate complexities in the assessment of information-integration category learning. *Cognition, 111*, 410-414.

Verguts, T., & Fias, W. (2009). Similarity and Rules United: Similarity- and Rule-Based Processing in a Single Neural Network. *Cognitive Science, 33*, 243-259.

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic Bulletin & Review, 8*, 168-176.

Wills, A. J., Lavric, A., Croft, G. S., & Hodgson, T. L. (2007). Predictive learning, prediction errors, and attention: Evidence from event-related potentials and eye tracking. *Journal of Cognitive Neuroscience, 19*, 843-854.

Wills, A. J., Lea, S. E. G., Leaver, L. A., Osthaus, B., Ryan, C. M. E., Suret, M. B., et al. (2009). A Comparative Analysis of the Categorization of Multidimensional Stimuli: I. Unidimensional Classification Does not Necessarily Imply Analytic Processing; Evidence From Pigeons (Columba livia), Squirrels (Scurius carolinensis), and Humans (Homo sapiens). *Journal of Comparative Psychology, 123*, 391-405.

Wills, A. J., Longmore, C.A. & Milton, F. (2011). Deliberative Processing and Overall Similarity in Free Classification. Manuscript submitted to *Journal of Experimental Psychology: Learning, Memory & Cognition.*

Wills, A. J. & Pothos, E. M. (2011). On the adequacy of current empirical evaluations of formal models of categorization. Manuscript submitted to *Psychological Bulletin.*

Winman, A., Wennerholm, P., Juslin, P., & Shanks, D. R. (2005). Evidence for rule-based processes in the inverse base-rate effect. *Quarterly Journal of Experimental Psychology, 58*, 789-815.

A.

Training			Test		
A+	B+	AB-	A?	B?	AB?
C-	D-	CD+	C?	D?	CD?
E+	F+	EF-	E?	F?	EF?
G-	H-	GH+	G?	H?	GH?
I+	J+		I?	J?	**IJ?**
		KL-	**K?**	**L?**	KL?
M-	N-		M?	N?	**MN?**
		OP+	**O?**	**P?**	OP?

B.

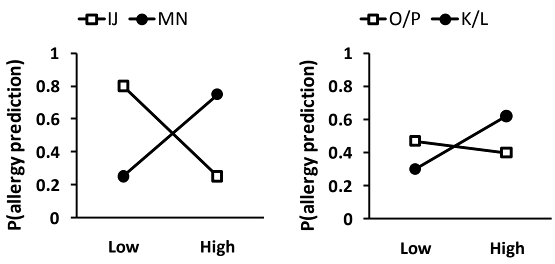

Figure 1. **A.** The training and test trial types in the Shanks and Darby (1998, Experiment 2) allergy prediction task; letters indicate foods eaten by a hypothetical patient Mr. X, + = patient develops an allergic reaction; - = patient does not develop an allergy reaction; ? = no feedback given. **B.** Critical test trials of Shanks and Darby (1998, Experiment 2) – probability of participants predicting an allergic reaction in Mr. X to novel meals, as a function of accuracy in the final block of the training phase. “Low” = low final training block accuracy (< 78% correct). “High” = high final training block accuracy (100% correct). Figure adapted from Shanks and Darby (1998).

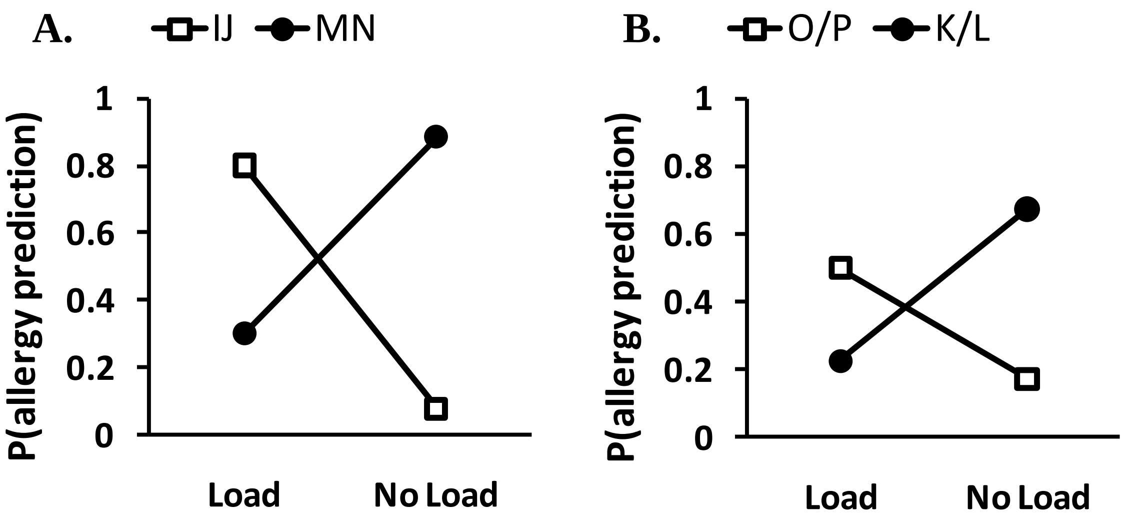

Figure 2. Results of Experiment 1; mean probability of participants predicting an allergic reaction in response to novel stimuli IJ, MN, O/P and K/L, as a function of the presence or absence of a concurrent load during training and test.

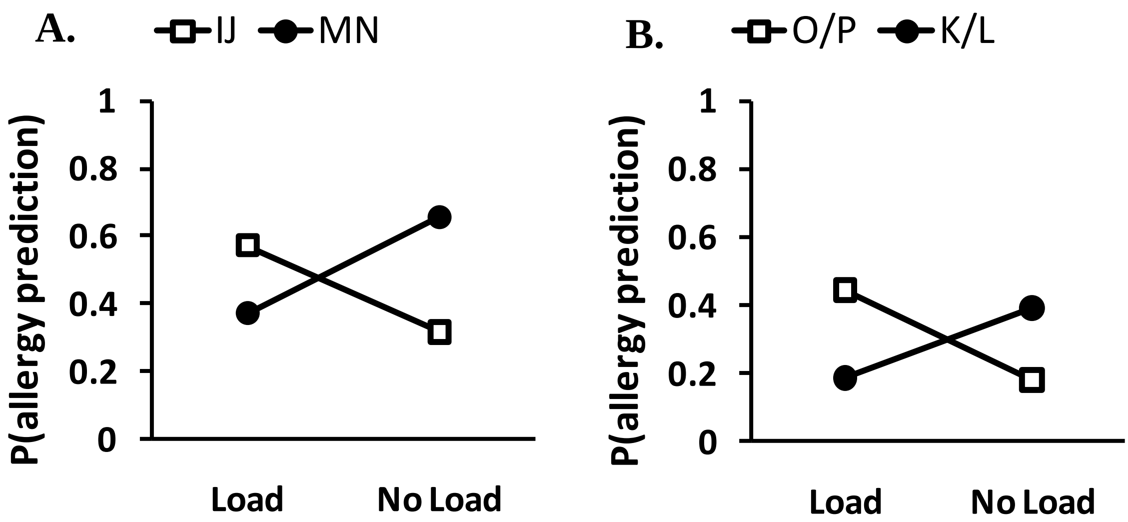

Figure 3: Results of Experiment 2; mean probability of participants predicting an allergic reaction in response to novel stimuli IJ, MN, O/P and K/L, as a function of the presence or absence of a concurrent load during training. The graphs average across conditions in which concurrent load was present and absent during the test phase, as the presence of test phase load had no significant effect on responding. The un-aggregated means (i.e. for each of the four conditions separately) are presented in the Supplementary Materials.

Supplementary Materials

Effects of Concurrent Load on Feature- and Rule-based Generalization in Human Contingency Learning

Andy J. Wills[1], Steven Graham[2], Zhisheng Koh[2], Ian P. L. McLaren[1], Matthew D. Rolland[1]

[1] University of Exeter

[2] National University of Singapore

In the main text, results of Experiment 2 were reported as averages across the *Test Load* factor, as the presence of concurrent load during test had no significant effect on any of our measures. For completeness, these Supplementary Materials report the same measures in an un-aggregated manner (i.e. separately for each of the four between-subject conditions).

	No Load	Train Load	Test Load	Both Load
Final training accuracy	.93 (.03)	.92 (.02)	.93 (.03)	.92 (.03)
Blocks to criterion	5.2 (1.9)	6.2 (1.3)	4.2 (1.6)	5.9 (1.4)

Table S1. Training phase results of Experiment 2, as a function of between-subjects condition. *Final training accuracy* is the mean proportion correct in the final training block. *Blocks to criterion* is the mean number of blocks completed in order to meet the training criterion. *No Load*, *Train Load*, *Test Load*, and *Both Load* indicate the four between-subject conditions of Experiment 2. Numbers in parenthesis are standard deviations. All results include only those participants who passed the training criterion.

	No Load	Train Load	Test Load	Both Load
IJ	.35 (.45)	.61 (.48)	.28 (.44)	.55 (.44)
MN	.57 (.49)	.35 (.49)	.74 (.41)	.40 (.45)
K/L	.32 (.42)	.19 (.34)	.46 (.44)	.19 (.32)
O/P	.24 (.34)	.47 (.38)	.12 (.23)	.42 (.33)
Familiar +ve	.90 (.11)	.89 (.11)	.91 (.08)	.89 (.10)
Familiar –ve	.11 (.10)	.14 (.11)	.08 (.11)	.10 (.10)

Table S2. Proportion of allergy responses in the test phase of Experiment 2, as a function of between-subjects condition and test item. K/L indicates the mean of responses to K and L; O/P correspondingly. Blue shading indicates a pair of compound test items for which the relative magnitude of allergy responses is consistent with opposites-rule generalization. Green shading = element test items consistent with opposites-rule. Salmon = compound test items consistent with feature-based generalization. Pink = element test items consistent with feature-based generalization. *Familiar +ve* is the mean of allergy responses to stimuli also presented in training and for which the correct response in training was “allergy”. *Familiar –*

ve is the corresponding mean for familiar stimuli for which the correct response was "no allergy". *No Load*, *Train Load*, *Test Load*, and *Both Load* indicate the four between-subject conditions of Experiment 2. Numbers in parenthesis are standard deviations. All results include only those participants who passed the training criterion.