

# The dimensionality of perceptual category learning: A state-trace analysis

**BEN R. NEWELL**

*University of New South Wales, Sydney, New South Wales, Australia*

**JOHN C. DUNN**

*University of Adelaide, Adelaide, South Australia, Australia*

AND

**MICHAEL KALISH**

*University of Louisiana, Lafayette, Louisiana*

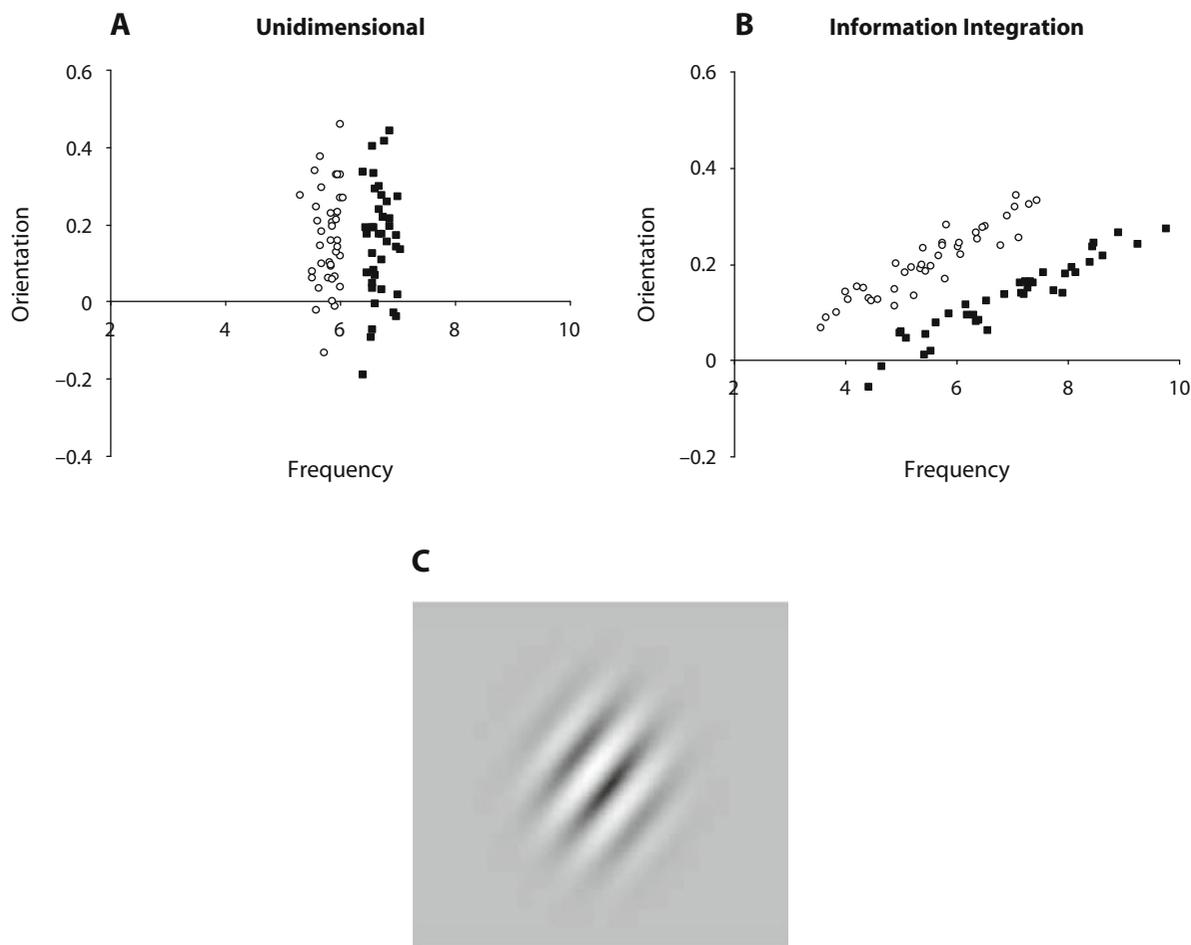
State-trace analysis was used to investigate the effect of concurrent working memory load on perceptual category learning. Initial reanalysis of Zeithamova and Maddox (2006, Experiment 1) revealed an apparently two-dimensional state-trace plot consistent with a dual-system interpretation of category learning. However, three modified replications of the original experiment found evidence of a single resource underlying the learning of both rule-based and information integration category structures. Follow-up analyses of the Zeithamova and Maddox data, restricted to only those participants who had learned the category task and performed the concurrent working memory task adequately, revealed a one-dimensional plot consistent with a single-resource interpretation and the results of the three new experiments. The results highlight the potential of state-trace analysis in furthering our understanding of the mechanisms underlying category learning.

The ability to categorize objects is a fundamental aspect of cognition. Given this central role, there is great interest in how people learn both natural and artificial categories. Research reveals that people show a high level of flexibility in the variety of category structures that they are able to learn (e.g., Ashby & Lee, 1991; Ashby & Maddox, 2005; Kruschke, 1992; Lagnado, Newell, Kahan, & Shanks, 2006; Medin & Schwanenflugel, 1981; Murphy, 2002; Nosofsky & Johansen, 2000; Nosofsky & Zaki, 1998). The goal of many theories and models of category learning is to account for this flexibility and variety.

A prominent theory of category learning that addresses this goal is the COVIS (competition between verbal and implicit systems) model proposed by Ashby, Alfonso-Reese, Turken, and Waldron (1998), which distinguishes between an explicit or verbal system and an alternative implicit system underlying learning. Much of the research testing COVIS has contrasted the learning of two distinct kinds of category structure: unidimensional (UD), or rule-based, and information integration (II). A set of multidimensional stimuli conform to a UD structure if they can be classified on the basis of a single, easily verbalized dimension. A similar set of stimuli conform to an II structure if they can be classified only on the basis of two or more different dimensions. This is shown schematically in Figure 1. Here, the filled squares and unfilled circles correspond to stimuli from two different categories. In

Figure 1A, the categories are defined by the level of one relevant dimension (spatial frequency of a perceptual stimulus). In Figure 1B, the categories are defined with respect to the levels of two relevant dimensions (spatial frequency and orientation). An example stimulus is shown in Figure 1C. COVIS proposes that UD structures are learned by the explicit or verbal system, whereas II structures are learned by the implicit system.

The verbal system of COVIS uses explicit rules, is dependent on working memory and executive attention (for storing and testing rules, respectively), and is suited to learning the easily verbalizable UD structures. The implicit system is not dependent on memory and attention; rather, it learns categories by learning the procedure for generating a response (i.e., the assignment of a category label). No verbalization is required (or perhaps even possible; see Ashby et al., 1998), thus suiting the system to the nonverbalizable II structures. The systems are also hypothesized to recruit different neuroanatomical structures. The verbal system relies on the anterior cingulate, the prefrontal cortices, and the head of the caudate nucleus (Ashby & Maddox, 2005; Filoteo et al., 2005; Love & Gureckis, 2007; Price, Filoteo, & Maddox, 2009; Seger & Cincotta, 2005, 2006). The implicit system is said to recruit the tail of the caudate nucleus (Ashby et al., 1998; Nomura & Reber, 2008; Seger & Cincotta, 2005).



**Figure 1.** (A) Distribution of stimuli in a unidimensional category structure. (B) Distribution of stimuli in an information integration category structure. (C) An example of a Gabor patch stimulus.

### Dissociations and the Appeal of State-Trace Analysis

The neuroscience evidence is one of the pillars of support for COVIS. However, evidence that different tasks are subserved by different brain regions does not, by itself, imply functionally distinct systems (Sherry & Schacter, 1987; see also Henson, 2006). Thus, behavioral data are crucially important for examining the potential interactions between multiple category learning systems. There is now a huge number of studies documenting *functional dissociations* between UD and II categorization tasks (see Ashby & Maddox, 2005; Maddox & Ashby, 2004, for reviews). In a typical experiment, a variable is found to affect learning of either the UD or II structure but to have little or no effect on learning the alternative structure. These studies have contributed significantly to our understanding of how different processes might combine during category learning. However, there are limits to the conclusions that can be drawn from studies employing this methodology (e.g., Dunn, 2003; Dunn & James, 2003; Dunn & Kirsner, 1988; Harley, Dillon, & Loftus, 2004; Loftus, 1978; Loftus, Oberg, & Dillon, 2004; Newell & Dunn, 2008). The crux of the argument is that dissociations are

neither necessary nor sufficient for the conclusion of multiple underlying processes (Newell & Dunn, 2008). This follows from the proposal that different tasks have different, and usually unknown, performance–resource functions (Norman & Bobrow, 1975; Shallice, 1988) that describe the relationship between an underlying cognitive resource (e.g., memory, perception, learning) and measurable performance on a given task. As a consequence, performance measures of these tasks are not commensurable, in the sense that they do not lie on a common scale and therefore cannot be directly compared. Since a dissociation consists of the observation of a significant difference in performance on one task across two or more experimental conditions coupled with the lack of a similar difference on a second task, it necessarily entails a comparison between a “large” difference on the first task and a “small” difference on the second; but, since the measures of performance on the two tasks are incommensurable, such a comparison has no meaning. To draw a parallel with the kind of stimulus illustrated in Figure 1, it would be analogous to the claim that a difference in orientation, measured in degrees, is larger or smaller than a difference in spatial frequency, measured in cycles per degree.

In the present context, the critical comparison is between relative differences in performance in learning II and UD structures across different experimental conditions. It cannot be assumed that these tasks have identical performance–resource functions because—if for no other reason—the relevant categories may differ in discriminability for the two structures (i.e., the average distance between members of Categories A and B may differ in II and UD structures; cf. Figures 1A and 1B). This implies that equivalent levels of learning may be revealed by different levels of performance, whereas different levels of learning may be revealed by equivalent levels of performance. Consequently, a change in the level of a single underlying resource may produce a relatively large change in UD learning coupled with a relatively small change (or no measurable change at all) in II learning, or vice versa. Such dissociations are therefore not necessary for drawing inferences about the number of processes or systems underlying category learning (Newell & Dunn, 2008).

In addition to being unnecessary to establish that different tasks depend upon more than one processing system, dissociations are also insufficient. We have proposed that *state-trace analysis* (Bamber, 1979) provides the appropriate basis for inferences of this kind (Newell & Dunn, 2008). State-trace analysis has already been successfully applied in a wide range of domains, including “remember–know” (RK) judgments in recognition memory (Dunn, 2008), judgments of learning (Jang & Nelson, 2005), contrast and visual memory (Harley et al., 2004), and the face-inversion effect (Loftus et al., 2004). The technique has the potential to illuminate the factors governing perceptual category learning by strengthening analyses of data and the conclusions that can be drawn from them. State-trace analysis is agnostic with regard to the nature or number of underlying processes, but is simply a tool that can be applied to determine the minimum number of processes that must be hypothesized to account for any given behavioral phenomenon.

State-trace analysis overcomes the problem of comparing different and unknown performance–resource functions and thereby allows inferences to be drawn concerning the number and organization of mental resources underlying multiple-task performance. An important analytic tool in this approach is the state-trace plot. This is simply a plot of performance on one task against performance on another task across different experimental conditions. The critical diagnostic feature of this plot concerns whether the data fall on a single, monotonically increasing curve, or otherwise. If the data fall on such a curve, there is no evidence of more than one underlying resource because, even though the performance–resource functions of each task may be unknown, it is reasonable to assume that they are both monotonically increasing; better performance is achieved as more of the underlying resource is available. It follows that, if the two tasks depend upon the same underlying resource (corresponding, in the present context, to a single category learning system), the relevant data should fall on a strictly increasing curve in the state-trace plot. If, on the other hand, more than one underlying resource is required, a two- (or multi-) dimensional state-trace plot

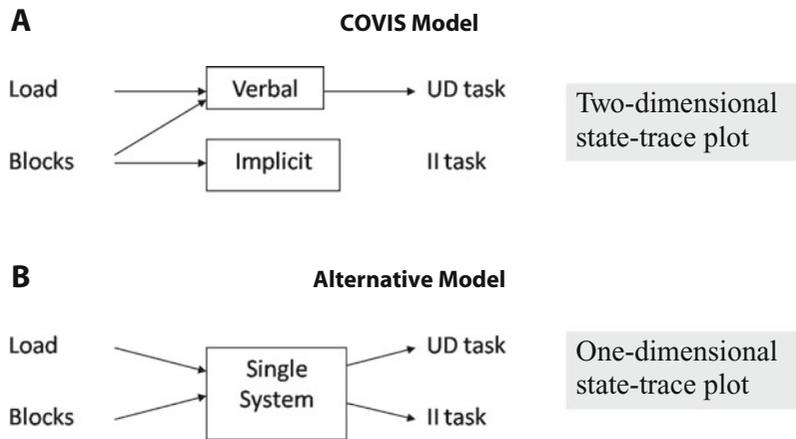
would be predicted, reflecting the differential involvement of two or more category learning systems on the tasks.

State-trace analysis also overcomes a second problem associated with the interpretation of dissociations. This concerns the fact that dissociations often rely on the failure to find a difference between two conditions, thereby continually running the risk of a Type II error. In fact, dissociation logic is the reverse of the normal approach to model testing. Using dissociation logic, the less complex one-dimensional model is rejected when a statistically significant effect is not found—that is, when evidence against the model has *not* been obtained. In contrast, in the normal approach to model testing, which includes state-trace analysis, the one-dimensional model is rejected only when a statistically significant effect is found—that is, when sufficient evidence against the model *has been* obtained. (See Appendix A for more details of the model testing procedure.)

In this article, we examine the potential involvement of multiple systems in perceptual category learning. First, we show that support for a multiple-system model depends on observing a two-dimensional state trace and that the existence, or otherwise, of a dissociation is orthogonal to this issue. We then apply state-trace analysis to just one of the many dissociations underpinning COVIS and examine the crucial question of whether the state trace is one- or two-dimensional. Our aim is to illustrate the potential of state-trace analysis for illuminating when multiple systems must be invoked to account for category learning.

### Examining the Impact of Cognitive Load: A Reanalysis of Zeithamova and Maddox (2006)

The dissociation we focus upon concerns the differential effect of working memory load on UD and II task performance recently reported by Zeithamova and Maddox (2006, hereafter Z&M).<sup>1</sup> Z&M predicted that the effect of a working-memory-demanding concurrent task would be greater on the learning of a UD task than of an II task. This prediction falls naturally from the COVIS framework because the system that learns the UD task (verbal) is reliant on working memory, whereas the system that learns the II task (implicit) is not. We chose to examine this dissociation because of the controversy surrounding the interpretation of cognitive load effects in category learning. There have been multiple claims and counterclaims concerning the reliance of II learning on working memory, with some authors even suggesting that higher working memory capacity can be detrimental for II learning. (For examples of these debates, see Ashby & Ell, 2002; DeCaro, Carlson, Thomas, & Beilock, 2009; DeCaro, Thomas, & Beilock, 2008; Foerde, Poldrack, & Knowlton, 2007; Newell, Lagnado, & Shanks, 2007; Nosofsky & Kruschke, 2002; Tharp & Pickering, 2009; Waldron & Ashby, 2001). One appealing aspect of the Z&M experiment is that, unlike many of the previous studies mentioned above, it used stimuli that do not easily lend themselves to verbalization (see Figure 1C), making it perhaps less likely that partial explicit knowledge could account for performance in the II task. The Z&M experiment thus provides perhaps the best evidence to date that cognitive load selectively affects



**Figure 2.** Two models of the effects of block (number of trials) and load (memory load) on performance in unidimensional (UD) and information integration (II) tasks. The COVIS model (A) predicts a two-dimensional state-trace plot, because both load and block affect the verbal learning system, whereas only block affects the implicit learning system. The alternative single-system model (B) predicts a one-dimensional state-trace plot, because the single system is affected by both variables.

UD learning, making it an excellent candidate for our re-examination with state-trace analysis.

Figure 2 provides a schematic summarizing the predictions for the effect of cognitive load. Figure 2A shows an “idealized” COVIS model in which blocks (number of trials) affects both the implicit and the verbal system but load (the presence of an additional working memory task) only impacts the verbal system and thus only affects learning of UD structures. In contrast, in a default single-system model shown in the lower panel (Figure 2B), this system is affected by both load and blocks and in turn determines performance on both the UD and the II tasks.

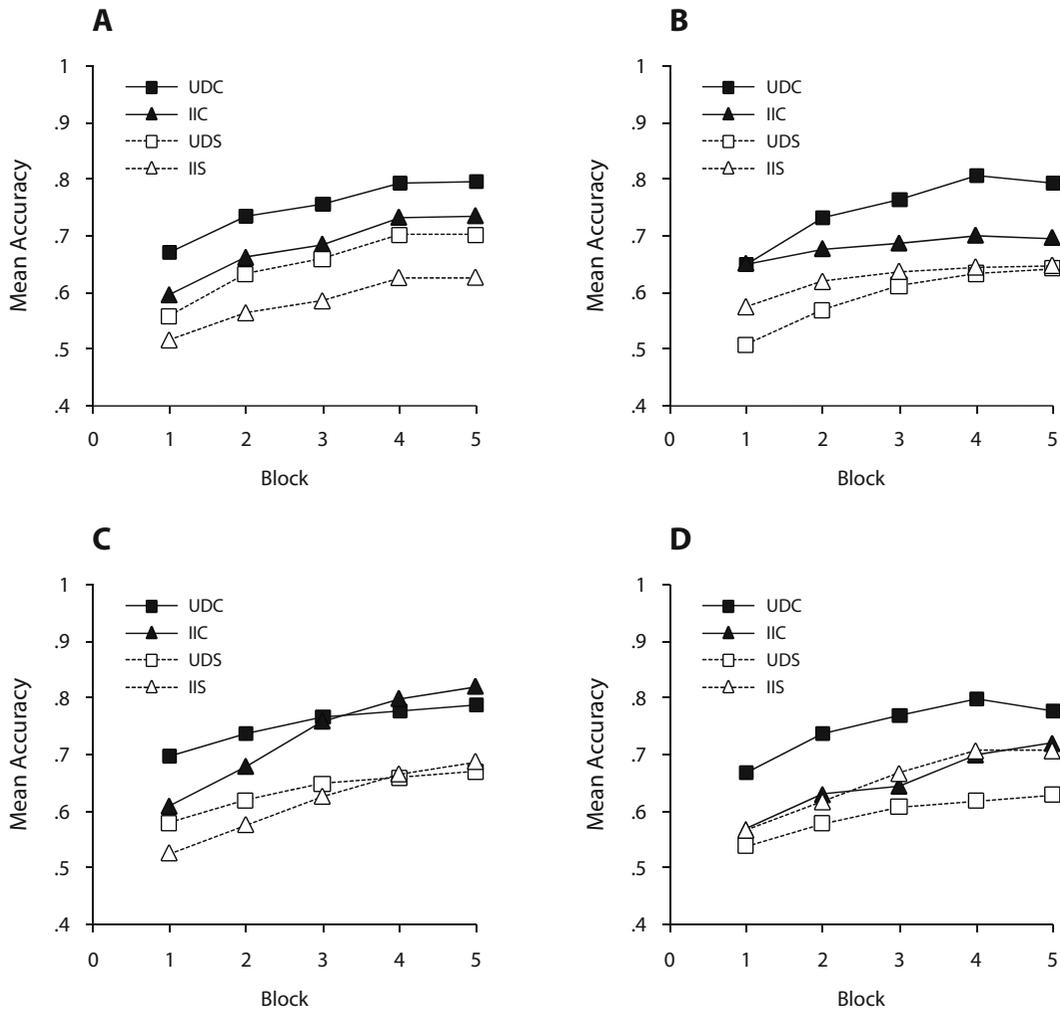
Z&M tested their prediction by giving four separate groups of participants either UD or II tasks, each with or without a concurrent memory-load task (a numerical “Stroop task”). The category learning task for all groups required participants to assign perceptual stimuli to the correct category, A or B. These stimuli varied on two dimensions—orientation of the lines and spatial frequency (perceived as bar width; see Figure 1C). For the UD task, the categorization decision could be made on the basis of one dimension (spatial frequency) and was thus verbalizable by a simple rule of the form “Respond A if the bars are thin and B if the bars are wide” (see Figure 1A). For the II task, both orientation and spatial frequency information needed to be integrated to make a correct decision, making it difficult to generate a simple, verbalizable rule (see Figure 1B). Given the structure depicted in Figure 1B, such a rule would need to be of the form “Respond A if the angle of the bars exceeds their width; otherwise respond B”; but, because angle and width are not commensurate, it is difficult if not impossible to implement such a rule (Stanton & Nosofsky, 2007).

### Comparison of One-Dimensional and Two-Dimensional State-Trace Plots

From the viewpoint of state-trace analysis, the crucial distinction between the two models shown in Figure 2

concerns their underlying dimensionality. According to the COVIS model shown in Figure 2A, there are two systems underlying the learning of UD and II structures and each may be differentially affected by different experimental variables, in this case by load and blocks. If II performance is plotted against UD performance, this should, in general, yield a two-dimensional state-trace plot (Newell & Dunn, 2008). In contrast, according to the alternative single-system model shown in Figure 2B, learning of UD and II structures depends upon the same underlying process or resource. No matter how two or more variables affect this resource, the resulting state-trace plot will always be one-dimensional and, by monotonicity, will trace out a monotonically increasing curve.

In order to demonstrate the implications of the two models for state-trace analysis, we constructed data sets based on the Z&M experiment but conforming either to the idealized COVIS model (Figure 2A) or the alternative single-system model (Figure 2B); Figures 3 and 4 show examples of the kinds of data sets that were produced. Figure 3 shows the results plotted as a function of training block, memory load, and task; Figure 4 shows their corresponding state-trace plots. Panels A and B in both figures show results derived from the single-system model. To construct these data, performance on UD and II tasks were modeled as different normal ogive functions of the same set of parameter values, representing a single level of learning in each condition. These functions, which varied in location and shape, were used to model the idea that performance on these two tasks was related to the underlying level of learning by a different performance–resource function. For the data shown in panel A, memory load affects both the UD and II tasks, approximately equally; there is no dissociation. For the data shown in panel B, memory load has a greater effect on the UD task than on the II tasks; there is a dissociation consistent with the predictions of COVIS. However, since the corresponding state-trace plot shows, this dissociation is completely



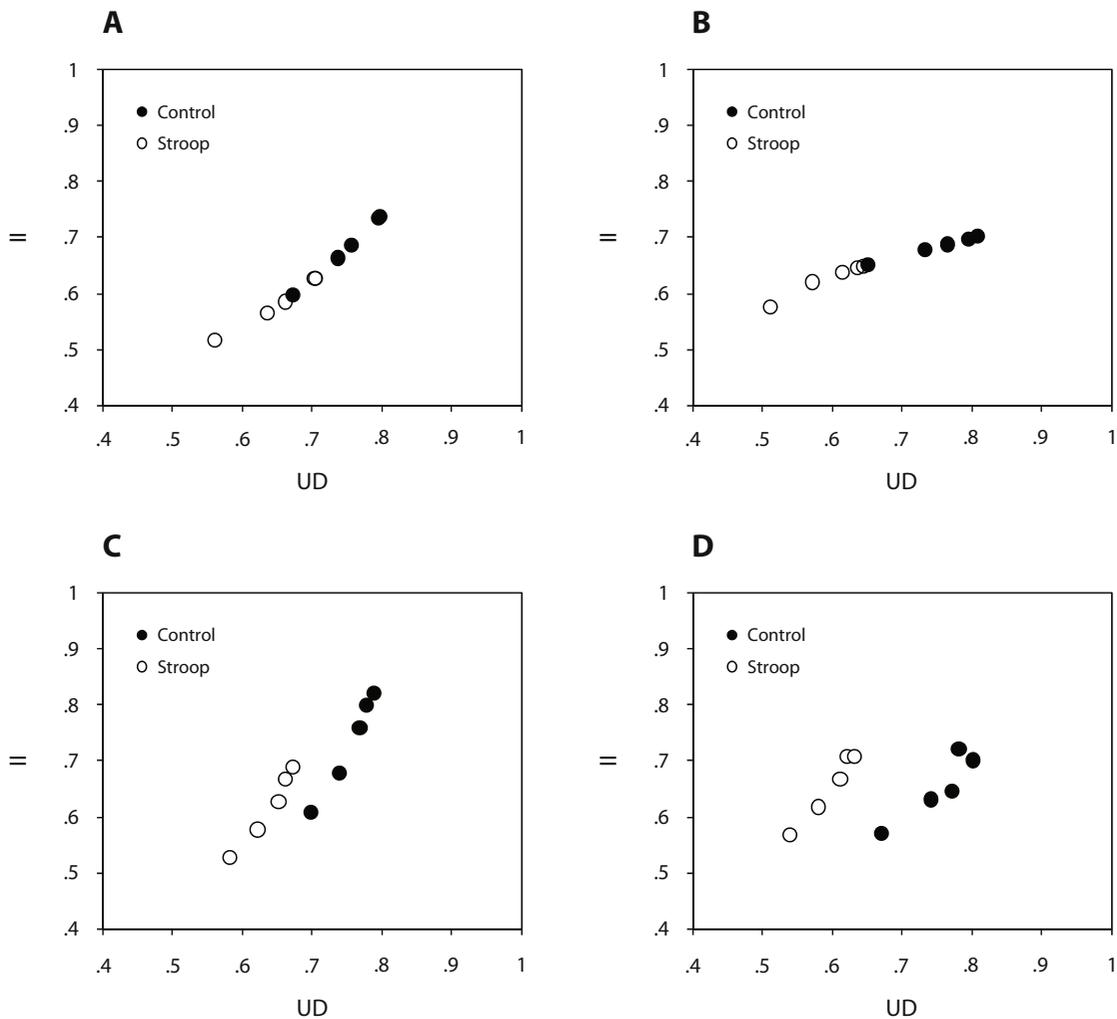
**Figure 3. Simulated data: Mean proportion correct for each group.** The top two panels show simulated data derived from a one-dimensional model, such as that shown in Figure 2B. The bottom two panels show simulated data derived from a two-dimensional model, such as COVIS (Figure 2A). (A) One-dimensional model, no dissociation present; load has equal effects on both UD and II tasks. The control conditions are denoted with solid lines and filled marks, the dual Stroop task conditions with broken lines and unfilled marks. (B) One-dimensional model, dissociation present; load has a greater effect on the UD task than on the II task. (C) Two-dimensional model, no dissociation present; load has equal effects on both UD and II tasks. (D) Two-dimensional model, dissociation present; load has a greater effect on the UD task than on the II task (consistent with the COVIS model). UDC, unidimensional rule-based control; IIC, information integration control; UDS, unidimensional rule-based Stroop; IIS, information integration Stroop.

consistent with a one-dimensional model, since all the data fall on a single monotonically increasing curve. This demonstrates that a dissociation is not sufficient to reject a single-system account.

The bottom two panels (C and D) in both figures show results derived from a multiple-system model. To construct these data, performance on UD and II tasks were modeled as different normal ogive functions of two sets of parameter values, representing the levels of learning of two underlying systems in each condition. For the data shown in Figure 3C, memory load affects both the UD and II tasks, approximately equally—there is no dissociation. Yet the corresponding state-trace plot is bidimensional—the data do not lie on a single monotonically increasing curve (Figure 4C). This demonstrates that a dissociation is not neces-

sary to reject a single-system account. Finally, for the data shown in Figure 3D, memory load has a greater effect on the UD task than on the II tasks; there is a dissociation consistent with the predictions of COVIS, and the corresponding state-trace plot is bidimensional (Figure 4D).

It is important to note that, although a bidimensional state-trace plot is consistent with a multiple-system model such as COVIS, it is also consistent with other sources of variability, or parameters, unrelated to the proposed learning systems. For example, bidimensionality may result if response criteria vary orthogonally to learning strength across conditions (Dunn, 2008). State-trace analysis therefore requires careful experimental control to eliminate unwanted or theoretically uninteresting sources of variability.



**Figure 4.** Simulated data: State-trace plots corresponding to the panels in Figure 3. (A) One-dimensional model, no dissociation present; load has equal effects on both UD and II tasks. The data fall on a monotonically increasing curve, with equal rates of change on both tasks. The control conditions are denoted with filled circles, the dual Stroop task conditions with unfilled circles. (B) One-dimensional model, dissociation present; load has a greater effect on the UD task than on the II task. The data fall on a monotonically increasing curve with greater rate of change on the UD task than on the II task. (C) Two-dimensional model, no dissociation present; load has equal effects on both UD and II tasks. The data fall on two curves displaced equally on both tasks. (D) Two-dimensional model, dissociation present; load has a greater effect on the UD task than on the II task (consistent with the COVIS model). The data fall on two curves displaced only on the UD task.

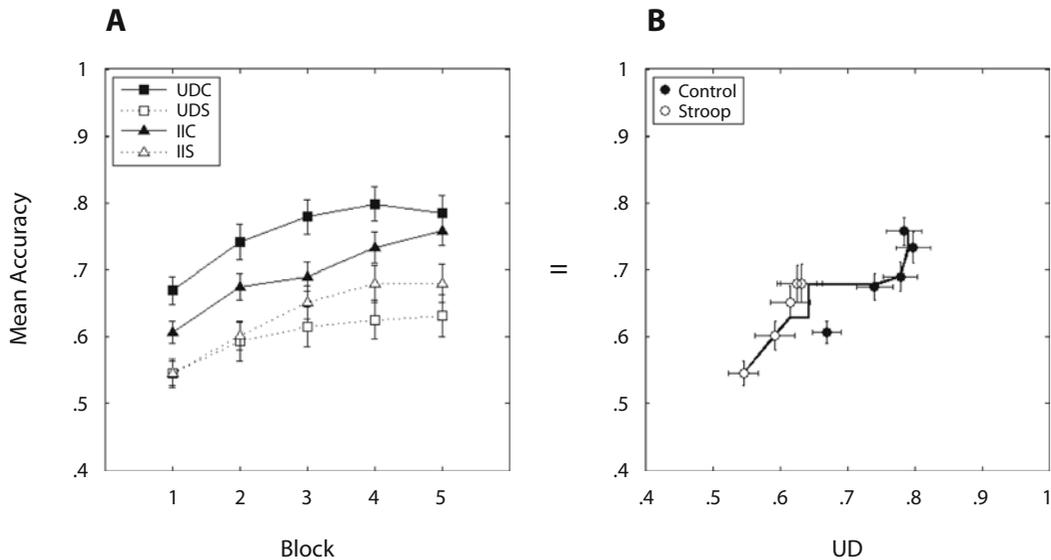
### State-Trace Analysis of Z&M

The example data shown in Figures 3 and 4 reveal that evidence that uniquely supports COVIS must satisfy two separate conditions: first that memory load differentially affects the UD task in contrast to the II task, and second that the resulting state-trace plot is not one-dimensional. We now turn to the question of whether the results found by Z&M fulfill these conditions.

Figure 5 shows the results found by Z&M and the corresponding state-trace plot.<sup>2</sup> COVIS predicts that the difference between performance on the control and Stroop conditions should be greater for participants learning the UD task than for those learning the II task. This pattern was duly found: The overall difference in accuracy across 400 trials between the UD Stroop and control groups was

15.6%, whereas the corresponding difference between the II groups was only 6.1%. An important feature of these data is that, although UD control performance is better than II control performance, UD Stroop performance is *worse* than the II Stroop performance.

As noted above, although the dissociation is consistent with COVIS, it does not by itself necessitate a multiple-system view. It is necessary to determine the dimensionality of the corresponding state-trace plot. This is shown in Figure 5B and reveals an apparent two-dimensional structure, consistent with COVIS. However, rather than relying on only a visual impression, we would like to be able to *test* whether the data require a two-dimensional solution. To do this, we have developed an inferential procedure specifically designed for the analysis of state-trace data.



**Figure 5.** (A) Mean proportion correct for each group in Zeithamova and Maddox’s (2006) Experiment 1, with .80 or above on the Stroop task only included ( $N = 142$ ). The control conditions are denoted with solid lines and filled marks, the dual Stroop task conditions with broken lines and unfilled marks. Error bars are standard errors of the mean. (B) State-trace plot of the data from panel A. The dark line shows the best-fitting monotonically increasing curve. UDC, unidimensional rule-based control; UDS, unidimensional rule-based Stroop; IIC, information integration control; IIS, information integration Stroop.

The procedure consists of two parts: a model fitting part, in which we fit a set of models to the data using maximum likelihood estimation (MLE); and a null hypothesis statistical testing (NHST) part, in which we use a Monte Carlo procedure to estimate the probability of the observed goodness of fit of each model given the null hypothesis that it is true.<sup>3</sup> Details of this method are given in Appendix A but can be summarized as follows. In the model fitting part, MLE is used to fit three different models to the data. These models are the *trace model*, the *one-dimensional model*, and the *nonoverlap model* (Heathcote, Brown, & Prince, 2009). The trace model implements a set of order constraints that we expect a priori to be true and is used to detect theoretically uninteresting bidimensionality. In the present case, this corresponds to the requirement that categorization performance improve across trials for each group. As can be seen in Figure 5A, there are some minor violations of this model between Blocks 4 and 5 for the UDC and IIS groups that are revealed as departures from monotonicity in the corresponding state-trace plot in Figure 5B. Such violations, if significant, are not evidence for the existence of multiple learning systems in accordance with the models presented in Figure 2.

If the trace model is accepted, the one-dimensional model is fit to the data. This model is nested within the trace model and specifies a set of additional constraints requiring that performance on the UD and II tasks be monotonic functions of each other across all the conditions of the experiment. Figure 5B shows the best-fitting monotonic function that is the result of fitting this model. Violations of this model correspond to negative associations, or “dips,” in the state-trace plot between any of the control conditions and any of the Stroop conditions. Such

violations are difficult to see in conventional plots of the data, as in Figure 5A, but are readily apparent in the state-trace plot (Figure 5B). In this case, it is clear that relatively large violations of the model occur between Blocks 3–5 of the Stroop condition and Block 1 of the control condition: the leftmost filled circle (data for Block 1 of the control condition) sits well below the three rightmost bunched unfilled circles (data for Blocks 3, 4, and 5 of the Stroop condition). There are also relatively small violations between Blocks 4 and 5 of the Stroop condition and Block 2 of the control condition: the second leftmost filled circle (data for Block 2 of the control condition) sits marginally below the two rightmost unfilled circles (data for Blocks 4 and 5 of the Stroop condition). The one-dimensional model tests to determine whether these violations are statistically significant.

If the one-dimensional model is accepted, the nonoverlap model is fit to the data. This model is nested within the one-dimensional model and specifies an ordering of conditions indicating nonoverlap of the data. This would occur if all the Stroop conditions were worse than all the control conditions on both tasks. In Figure 5B, it can be seen that violation of this model depends primarily on the data from Block 1 of the control condition. If the level of II performance happened to be higher—at the same level as Blocks 4 and 5 in the Stroop condition—there would be no overlap in the data. The reason for testing nonoverlap is that it is trivial to fit a monotonically increasing function to nonoverlapping data. As a result, acceptance of the nonoverlap model tells us that, although the one-dimensional model may fit the data, it does so for theoretically uninteresting reasons; essentially, there has been a failure in experimental design.

In summary, evidence that a state trace is bidimensional for theoretically interesting reasons requires that the trace model be accepted and the one-dimensional model be rejected. Evidence that a state trace is unidimensional for theoretically interesting reasons requires that the trace model and the one-dimensional model be accepted and that the nonoverlap model be rejected (see Appendix A for a description of how  $p$  values for these models are estimated).

We applied the MLE/NHST method to the Z&M data. The resulting best-fitting monotonic function is shown by the dark line in Figure 5B. Goodness of fit of this and the trace and nonoverlap models is measured by the  $G^2$  statistic. Since these models are nested within each other, we report and test the *difference* in  $G^2$ ,  $\Delta G^2$ , between each model and that in which it is nested. Thus the trace model is compared with an unrestricted model, for which  $G^2 = 0$  necessarily, the one-dimensional model is compared with the trace model, and the nonoverlap model is compared with the one-dimensional model. For the trace model,  $\Delta G^2 = 0.13$ ,  $p = .986$ ; for the one-dimensional model,  $\Delta G^2 = 2.43$ ,  $p = .658$ ; and for the nonoverlap model,  $\Delta G^2 = 4.81$ ,  $p = .352$ . Whereas this pattern is consistent with a unidimensional model, because it was not possible to reject the nonoverlap model, the data are actually equivocal.

### Interim Summary

The state-trace analysis and the standard analysis of the Z&M data highlight the differing logic of the two approaches. According to the state-trace analysis, the failure to reject the one-dimensional model leads us to conclude that there is no statistical evidence for bidimensionality, therefore no evidence for the involvement of multiple learning resources. According to the standard analysis, the failure to reject the hypothesis of a difference between two conditions would lead us, as it led Z&M, to conclude that there *is* statistical evidence for bidimensionality and hence evidence *for* the involvement of multiple learning resources. We noted earlier the controversy surrounding the interpretation of cognitive load effects on category learning performance (e.g., Ashby & Ell, 2002; Newell et al., 2007; Nosofsky & Kruschke, 2002). Given this controversy and the marginal nonoverlap in the data, we wanted to be sure that the reinterpretation of Z&M suggested by the state-trace analysis was justified. For this reason, we ran three new experiments, all modified replications of Z&M Experiment 1.

## EXPERIMENT 1

Experiment 1 followed the same procedure as Z&M's Experiment 1. The stimuli used by Z&M differ in terms of spatial frequency and line orientation (see Figure 1C) but also in terms of the *discriminability* ( $d'$ ) of stimuli assigned to the different categories (i.e., the average distance between items from Categories A and B; cf. Figures 1A and 1B). We sampled from the same parameters used by Z&M (see Table 1) and followed Z&M by adjusting the distributions to produce greater discriminability in the II

**Table 1**  
Category Distribution Parameters for the Unidimensional and Information Integration Category Structures Used in Experiments 1–3 and in Zeithamova and Maddox (2006, Experiment 1)

Category Structure	$\mu_x$	$\mu_y$	$\sigma_x^2$	$\sigma_y^2$	cov <sub>xy</sub>
Unidimensional					
Category A	280	125	75	9,000	0
Category B	320	125	75	9,000	0
Information Integration					
Category A	268	157	4,538	4,538	4,351
Category B	332	93	4,538	4,538	4,351

condition in order to attempt to equalize difficulty in the control conditions. Our resulting  $d'$  estimates were 4.3 for UD and 6.7 for II.

### Method

**Participants.** One hundred thirty-five undergraduate students from the University of Adelaide participated in return for course credit or a payment of AU\$12 each. Each participant completed one experimental condition, with 32 participants in unidimensional control (UDC), 35 in information integration control (IIC), and 34 participants respectively in unidimensional Stroop (UDS) and information integration Stroop (IIS).

**Stimuli and Apparatus.** The categorization stimuli were generated using the same procedures as Z&M. The stimuli were Gabor patches with varying spatial frequency and spatial orientation. Forty Category A stimuli and 40 Category B stimuli for the UD categories were generated by sampling randomly from the same two bivariate normal distributions used by Z&M. The parameters used are shown in Table 1. The 80 stimuli for the II categories were obtained by rotating the 80 unidimensional stimuli clockwise by 45° around the center of the spatial-frequency–spatial-orientation space, then shifting the spatial frequency and spatial orientation to achieve an appropriate level of discriminability ( $d'$ ). The Gabor patch stimuli were then produced using MATLAB (MathWorks, Natick, MA) routines from the Psychophysics Toolbox (Brainard, 1997). Each stimulus was 200 × 200 pixels and was centered on the computer screen. The numerical “Stroop” task used in the two dual-task conditions sampled without replacement from the range 2–8. On 85% of trials the numerically larger number was physically smaller (95 vs. 180 pixels tall). All stimuli were presented on a gray background.

**Procedure.** The experiment consisted of five 80-trial blocks of trials. Within each block, all 80 stimuli were presented in a random order (with different orders for each participant). Participants in the two control conditions (UDC and IIC) were told to learn via corrective feedback which of two categories (A or B) each stimulus belonged to. On each trial, a single Gabor patch stimulus remained on screen until the participant responded by pressing either the “Z” button or the “?” button on the computer keyboard. Feedback, provided for 1,000 msec, consisted of a low tone for an incorrect response and a high tone for a correct one. Following feedback were a 1,000-msec delay and a 1,000-msec intertrial interval (ITI). Participants in the two dual-task “Stroop” conditions were presented with both the categorization stimulus (Gabor patch) and the two Stroop numbers concurrently. The numbers appeared to the right and left of the Gabor patch for 200 msec and were then replaced for 200 msec by a white rectangular mask. The participant first responded to the categorization stimulus (by pressing “Z” or “?”); then, after 1,000 msec of feedback, and a delay of 1,000 msec, the word “value” or “size” appeared on the screen. The participant then indicated which side of the screen the number with the larger value or size had appeared. The response was followed by 1,000 msec of corrective feedback and the 1,000-msec ITI. The timing of each trial was identical to that used in Z&M.

**Table 2**  
**Percentage of Participants Included in the Filtered Data Analyses of**  
**Experiments 1–3 and in Zeithamova and Maddox (2006, Experiment 1)**

Experimental Condition	Z&M (2006)		New Experiments					
	Experiment 1 <sup>††</sup>		Experiment 1 <sup>††</sup>		Experiment 2 <sup>††</sup>		Experiment 3 <sup>††</sup>	
	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%
UDC	32/41	78	31/32	97	31/32	97	31/32	97
UDS <sup>†</sup>	12/35	34	19/26	73	42/55	76	20/32	62
IIC	30/34	88	28/35	80	28/35	80	28/35	80
IIS <sup>†</sup>	19/32	59	8/23	34	32/62	52	17/29	58

Note—*Filtered data* refers to participants with 80% concurrent task performance and 65% final block accuracy in the category learning task. UDC, unidimensional control; UDS, unidimensional Stroop; IIC, information integration control; IIS, information integration Stroop. The control groups are the same for all three of the new experiments. <sup>†</sup>See text for details of the Stroop/concurrent tasks used in each experiment. <sup>††</sup>For the frequencies, the denominator in each condition is the number of *unfiltered* participants and the numerator is the number of *filtered* participants. *Unfiltered data* refers to all participants scoring above 80% on the concurrent task. Exclusions on the basis of this concurrent task criterion were: 28 participants in Z&M Experiment 1; and 19, 1, and 7 participants in Experiments 1–3, respectively. None of these excluded participants was included in any of our (or Z&M's) analyses.

## Results and Discussion

**Data exclusion and filtering.** We followed Z&M by first excluding participants performing below 80% correct in the Stroop task. We then filtered the data for category learning accuracy by following the practice undertaken in many investigations of COVIS (but not by Z&M—an issue we return to later) of defining “learners” as participants achieving accuracy of 65% or above in the final block of trials (e.g., Maddox & Ing, 2005; Zeithamova & Maddox, 2007). (For a block of 80 trials, .65 approximates the value of the binomial probability distribution for an alpha value of .05). Thus, *unfiltered* data refers to all participants (except those performing below 80% in the Stroop task), and the *filtered* data refers to participants who performed above 80% correct in the Stroop task and achieved accuracy of 65% or more in the final block of trials.

To foreshadow our results, we did not find major differences between filtered and unfiltered data sets, although we focused on those participants who surpassed both performance criteria (*filtered data*) because our interest is in how participants *learn categories under increased cognitive load*. Any noteworthy differences between these data and the unfiltered data are described in the Notes, and figures displaying the unfiltered data for each experiment are presented in Appendix B. Table 2 shows the impact of filtering data on each condition.

**Stroop task performance.** Mean Stroop task accuracy was .87 ( $SD = .03$ ) in the IIS condition and .90 ( $SD = .05$ ) in the UDS condition. There was no significant difference in level of accuracy between the groups [ $F(1,25) = 2.75$ ,  $p = .11$ ], suggesting that effort and cognitive resources allocated to the Stroop task were equal in both conditions.

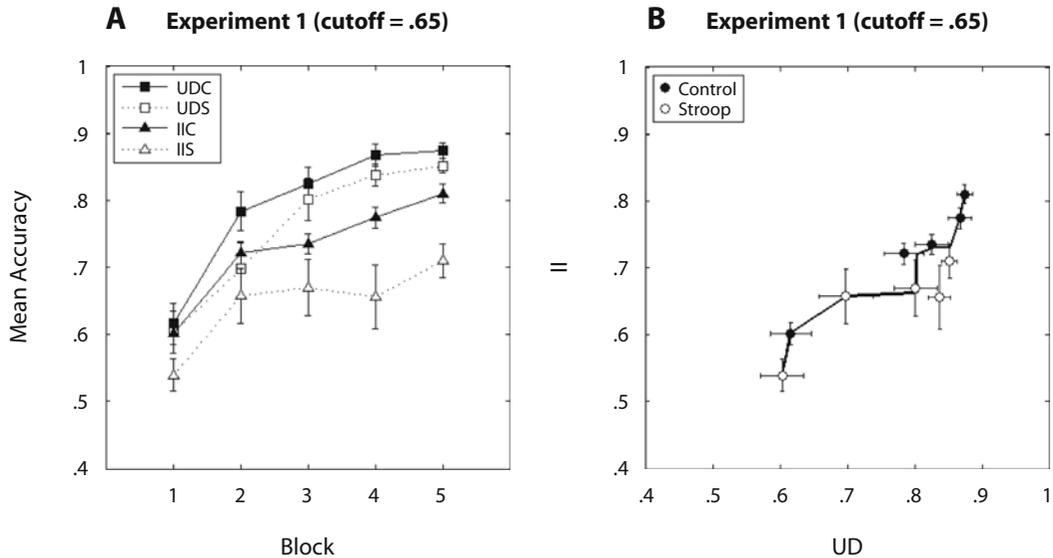
**Category learning performance.** For each participant, we computed the mean proportion correct for each block of 80 trials. These data are shown in Figure 6A. The figure shows that participants in all conditions improve across blocks, that performance in the UD conditions is better than that in the II conditions, and that there is a detrimental effect of load in both category structures. Statistical analyses confirmed these impressions. A 2 (load: Stroop vs. control)  $\times$  2 (category structure: II vs. UD)  $\times$

5 (Blocks 1–5) mixed model ANOVA revealed a main effect of load [ $F(1,82) = 7.23$ ,  $p = .009$ ], a main effect of category structure [ $F(1,82) = 16.58$ ,  $p < .0001$ ], and a main effect of block [ $F(4,328) = 75.74$ ,  $p = .001$ ]. Block also interacted significantly with category structure [ $F(4,328) = 4.35$ ,  $p = .002$ ], reflecting the steeper learning curves for the UD over the II conditions. The interaction between load and category structure was not significant [ $F(1,82) = 1.17$ ,  $p = .282$ ], although the figure indicates that there is a *larger* numerical effect of load on II than on UD; this is opposite to the prediction of the COVIS model. Figure 6B shows the state-trace plot of the filtered data and little evidence of bidimensionality. This is confirmed by the statistical analysis.<sup>4</sup> The trace model fits these data almost perfectly ( $\Delta G^2 = 0.10$ ,  $p = .996$ ), and the one-dimensional model cannot be rejected ( $\Delta G^2 = 2.52$ ,  $p = .243$ ), whereas the nonoverlap model clearly fails ( $\Delta G^2 = 63.55$ ,  $p < .001$ ).

The results clearly indicate that when those participants who both learned the category task and attended to the concurrent task are analyzed, both the standard and the state-trace analyses lend no support to the claim that load has differential effects on the learning of II and UD structures. Thus, Experiment 1 supports the “striking” conclusion from our reanalysis of Z&M that a single-resource model cannot be rejected.

## EXPERIMENT 2

One problem with filtering the data from Experiment 1 was the high percentage of exclusions from the IIS group; after filtering, only 34% (8/23) of participants remained (see Table 2). It appears that too many of the participants in the IIS condition, perhaps finding the task too difficult and thus not motivating, simply “gave up,” making meaningful comparisons of this condition with the others problematic. In Experiment 2, we attempted to increase motivation in the category learning task by reducing the attention that needed to be paid to execute the concurrent task successfully. To reduce effort, the Stroop task was modified so that participants had only to answer a question about the *value* of the numbers that appeared on



**Figure 6.** (A) Mean proportion correct for all conditions from Experiment 1; only participants scoring .65 or above on the final block of the category task and .80 or above on the Stroop task are included ( $N = 86$ ). The control groups are denoted with solid lines and filled marks, the dual Stroop task conditions with broken lines and unfilled marks. Error bars are standard errors of the mean. (B) State-trace plot of the data from panel A. The dark line shows the best-fitting monotonically increasing curve. UDC, unidimensional rule-based control; UDS, unidimensional rule-based Stroop; IIC, information integration control; IIS, information integration Stroop.

either side of the Gabor patch stimulus. This requirement lessened the load on working memory, because participants needed only to encode one attribute (value) of the presented numbers rather than both (size and value).

## Method

**Participants.** One hundred eighteen undergraduates from the University of Adelaide participated in Experiment 2 in return for course credit or a payment of AU\$12 each. Sixty-three participants completed the information integration simple-Stroop condition (II-SS), and 55 completed the unidimensional simple-Stroop condition (UD-SS). New control conditions were not run; in the Results section, we compare performance on the two new dual-task conditions with the relevant controls from Experiment 1.

**Stimuli, Apparatus, and Procedure.** The experiment was identical to Experiment 1 in all aspects except the requirements for the Stroop task. In contrast to Experiment 1, participants only had to make a response about the *value* of a number appearing on the left or right of the Gabor patch stimulus.

## Results and Discussion

Table 2 shows that the aim of reducing participant attrition after filtering was achieved. In both load conditions, there was an increase in the number of participants remaining after filtering the data, with the more marked improvement in the II condition. All analyses are based on the filtered data.

**Stroop task performance.** Mean Stroop task accuracy was .97 ( $SD = .03$ ) in the II-SS condition, and .98 ( $SD = .03$ ) in the UD-SS condition, and did not differ significantly between groups [ $F(1,73) = 1.10, p = .29$ ].

**Category learning performance.** Figure 7A shows the data from the UD-SS and the II-SS groups plotted alongside the IIC and UDC groups from Experiment 1. The pattern is very similar to that shown in Figure 6A and there

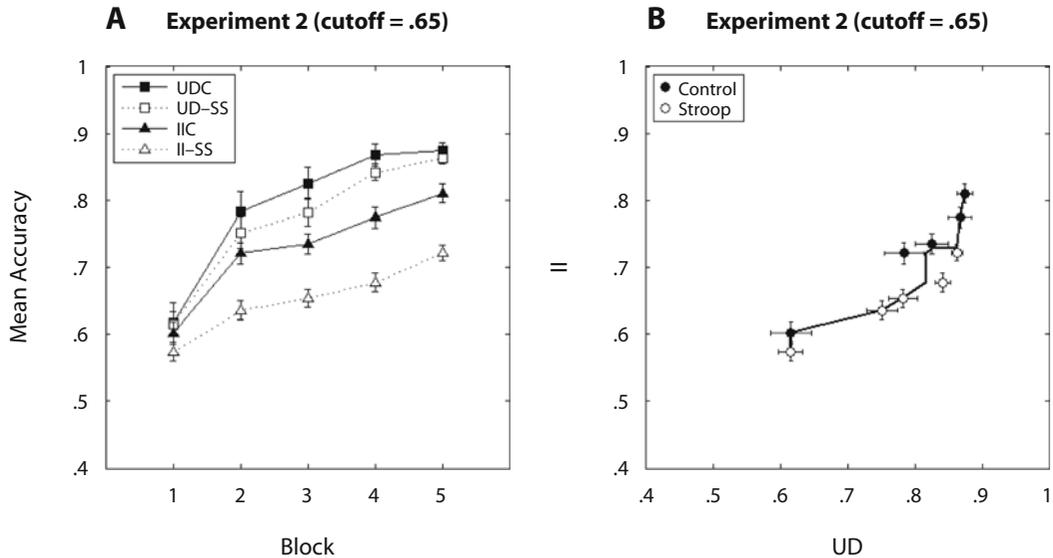
is no suggestion of the crossover interaction between load and category structure. The ANOVA revealed main effects of load [ $F(1,129) = 12.99, p < .001$ ], category structure [ $F(1,129) = 44.94, p < .001$ ], and a marginally significant interaction between category structure and load [ $F(1,129) = 3.83, p = .053$ ]; this interaction is due to load having a larger effect on the II structure than on UD. This is opposite to the prediction of COVIS. There was also a main effect of block [ $F(4,516) = 150.1, p < .001$ ], and block interacted significantly with load [ $F(4,516) = 7.93, p < .001$ ].

Figure 7B shows the state-trace plot of the filtered data. In this case, the trace model fits perfectly ( $\Delta G^2 = 0$ ), and although there is some evidence of bidimensionality, it is not possible to reject the one-dimensional model ( $\Delta G^2 = 4.59, p = .085$ ), whereas the nonoverlap model is clearly rejected ( $\Delta G^2 = 113.48, p < .001$ ). It should be noted that to the extent that the data are bidimensional, it is in the *opposite* direction to that predicted by COVIS. These data suggest that if load has a differential effect on categorization performance, it has a greater effect on the II structure than on the UD structure.<sup>5</sup>

The results of Experiment 2 contribute to the increasingly consistent picture and lend further weight to the reinterpretation of the Z&M data: A single underlying resource is sufficient to explain learning of II and UD structures in the presence and absence of additional cognitive load.

## EXPERIMENT 3

Experiments 1 and 2 showed either similar effects of load on learning II and UD category structures or a numerically larger effect on the II structure. In a final attempt to find the predicted bidimensionality in the data, Experi-



**Figure 7.** (A) Mean proportion correct for the simple-Stroop conditions of Experiment 2 and the control conditions from Experiment 1; only participants scoring .65 or above on the final block of the category task and .80 or above on the Stroop task are included ( $N = 133$ ). The control groups are denoted with solid lines and filled marks, the dual simple-Stroop task conditions with broken lines and unfilled marks. Error bars are standard errors of the mean. (B) State-trace plot of the data from panel A. The dark line shows the best-fitting monotonically increasing curve. UDC, unidimensional rule-based control; UD-SS, unidimensional rule-based simple Stroop; IIC, information integration control; II-SS, information integration simple Stroop.

ment 3 used an entirely different concurrent task. Z&M chose the numerical Stroop task for their load conditions because the standard Stroop task has been shown to require working memory and selective attention and because it strongly activates “the anterior cingulate and prefrontal cortex . . . , neural structures associated with the explicit-hypothesis testing system, but not with the implicit procedural learning system proposed in COVIS” (Z&M, p. 389). In Experiment 3, we chose an alternative concurrent task that fulfills the same criteria. This task required participants to maintain a five-digit string, presented before the category stimulus, then recall the location of a specific digit after responding to and receiving feedback on the category task. Jameson, Hinson, and Whitney (2004) argued that this task impairs decision making by placing demands on the central executive processes of working memory. Such processes are claimed to involve the prefrontal cortex and anterior cingulate (Baddeley, 2003), thus implying that the digit maintenance and recall task serves the required purpose of impacting the processes and structures thought to underlie the verbal system of COVIS.

## Method

**Participants.** Sixty-eight undergraduate students from the University of New South Wales participated in the experiment in return for course credit. Thirty-four participants completed the unidimensional digit recall condition (UD-DR) and 34 completed the information integration digit recall condition (II-DR). New control conditions were not run; in the Results section, we compare performance on the two new dual-task conditions with the relevant controls from Experiment 1.

**Stimuli, Apparatus, and Procedure.** The experiment was identical to Experiment 1 in all aspects except the requirements for the dual task. On each trial, participants were presented a string of

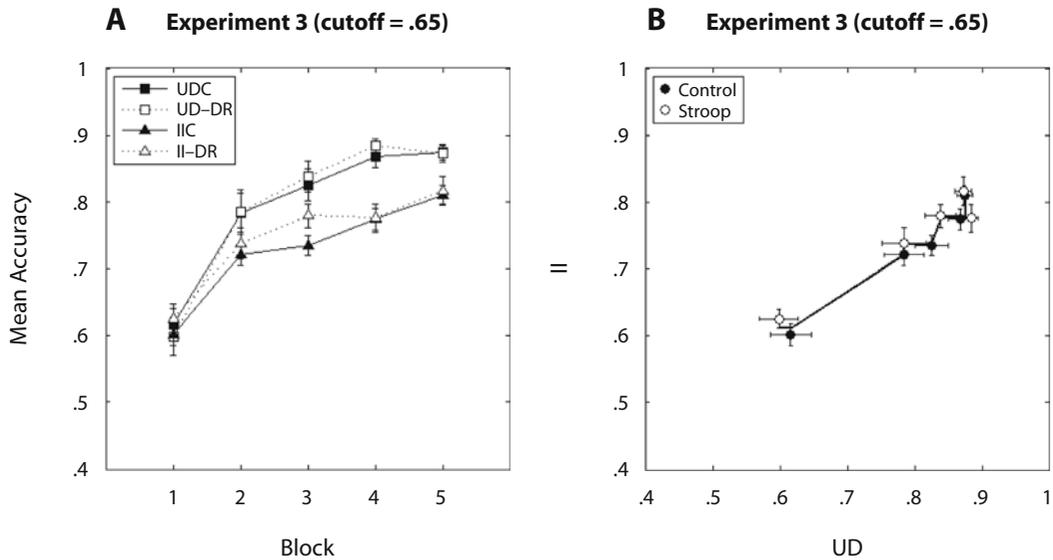
five digits composed of the numbers 1–5, randomly arranged (e.g., “25341”). The string remained onscreen for 2,000 msec, and participants were asked to try to remember the identity and position of each component digit. The digit string was then replaced by the Gabor patch stimulus, which remained onscreen until a categorization response was made. Following corrective feedback (high/low tone) and a 1,000-msec delay, a single “probe” digit from the original string was presented onscreen and participants were asked to recall which digit had been presented directly to the right of the probe. In the example above, if a “3” had been presented as a probe, the correct answer would have been “4.” Participants responded via the keyboard, and corrective feedback (high/low tone) followed. Finally, there was a 1,000-msec ITI.

## Results and Discussion

The impact of filtering participants on the basis of digit recall and category task performance is shown in Table 2. In all conditions, 50% or more of the participants remained after filtering. All analyses are based on the filtered data.

**Digit recall performance.** Mean digit recall performance was .95 ( $SD = .04$ ) in the II-DR group and .96 ( $SD = .03$ ) in the UD-DR group. Accuracy did not differ significantly between the groups [ $F(1,35) = 0.709$ ,  $p = .40$ ].

**Category learning performance.** Figure 8A displays the data from the two digit recall load conditions alongside the control conditions from Experiment 1. The figure suggests that load had very little effect on performance. A 2 (load: digit recall vs. control)  $\times$  2 (category structure: II vs. UD)  $\times$  5 (Blocks 1–5) mixed model ANOVA confirmed this impression, revealing a main effect of category structure [ $F(1,92) = 11.43$ ,  $p < .0001$ ], and a main effect of block [ $F(4,524) = 101.98$ ,  $p < .001$ ], but no effect of load ( $p > .50$ ). Block interacted significantly



**Figure 8.** (A) Mean proportion correct for the dual-task digit recall conditions of Experiment 3 and the control conditions from Experiment 1; only participants scoring .65 or above on the final block of the category task and .80 or above on the digit recall task are included ( $N = 96$ ). The control groups are denoted with solid lines and filled marks, the dual digit recall task conditions with broken lines and unfilled marks. Error bars are standard errors of the mean. (B) State-trace plot of the data from panel A. The dark line shows the best-fitting monotonically increasing curve. UDC, unidimensional rule-based control; UD-DR, unidimensional rule-based digit recall; IIC, information integration control; II-DR, information integration digit recall.

with category structure [ $F(4,368) = 6.07, p < .0001$ ], reflecting a more rapid improvement for the UD conditions. Figure 8B shows the state-trace plot of the filtered data. Once again, the data are consistent with both the trace and one-dimensional models<sup>6</sup> ( $\Delta G^2 = 0.10$  and  $\Delta G^2 = 0.31$ , respectively; both  $ps$  n.s.), with no indication of a failure of overlap ( $\Delta G^2 = 179.68, p < .001$ ).

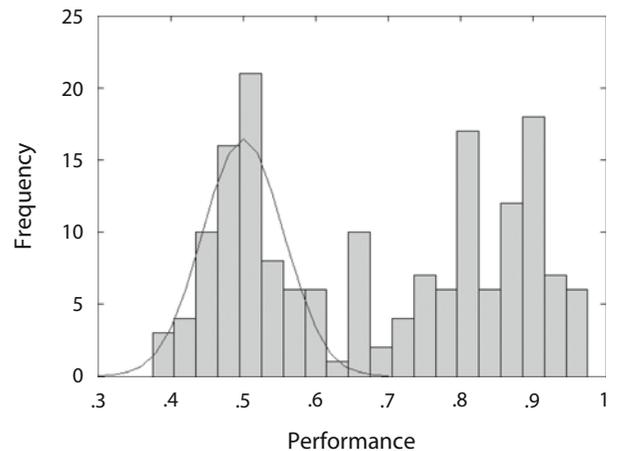
**RECONCILING OUR DATA WITH Z&M (EXPERIMENT 1)**

Our aim in running the three new experiments was to increase our confidence that the reinterpretation of the Z&M data suggested by state-trace analysis was not specific to their data. In three modified replications, we found evidence that a single-resource model could not be rejected. This pattern was clear in our data, but it appeared very different to the pattern found by Z&M, even though this too was formally consistent with the one-dimensional model. In this section, we examine the reasons for this difference and show how these can be understood using the principles of state-trace analysis.

A key difference between our analyses and Z&M is that they did not filter their data. Despite reporting a bimodal distribution of scores for participants in the UDC condition, Z&M did not apply the 65% learning accuracy criterion to the data from their Experiment 1. Figure 9 is a histogram showing the distribution of categorization accuracy scores (proportion correct) from all participants for the final block of 80 trials from Z&M Experiment 1, together with the binomial probability associated with pure chance responding. These data are collapsed across the four conditions (UDC, IIC, UDS, IIS) and indicate a

clear “break point” at .65, which is also the  $p = .05$  value for rejecting the hypothesis that responding was due to chance. This suggests that, although Z&M did not apply the 65% accuracy criterion to their data, there is a priori evidence that its application might have been warranted.

The inclusion of a significant proportion of nonlearners in the data can be sufficient to induce the apparent bidimensionality found by Z&M. This arises because variation in the proportion of nonlearners across the conditions



**Figure 9.** A histogram showing the distribution of categorization accuracies (proportion correct) for the final block of trials of Zeithamova and Maddox (2006), Experiment 1. There is a clear division at .65, providing a rationale for classifying participants scoring above .65 as “learners” and those below .65 as “nonlearners.” The curve shows the binomial probability distribution generated by chance performance.

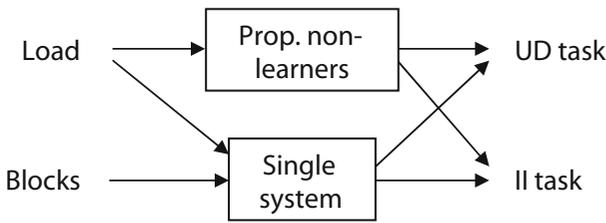


Figure 10. Schematic diagram of the effect of the proportion of nonlearners on mean performance on UD and II tasks.

of the experiment may have the same effect as that produced by multiple learning systems. To understand this, we first write a general equation for the unidimensional model corresponding to Figure 2B; that is,

$$UD = f(x) \text{ and } II = g(x),$$

where  $x$  is a parameter representing level of learning and a function of both cognitive load and the number of trials, and  $f(\cdot)$  and  $g(\cdot)$  are arbitrary monotonic functions that map  $x$  onto the observed levels of performance on the UD and II tasks, respectively. Suppose now that a proportion of participants fail to learn, performing at chance throughout the experiment, and that this proportion is a function of load, which varies in potentially different ways for the UD and II groups. These proportions can be described in functional form as  $p(\text{load})$  for the UD task, and  $q(\text{load})$  for the II task. Performance then becomes a mixture of two different learning profiles; that is,

$$UD = f(x) \cdot p(\text{load}) + 0.5[1 - p(\text{load})],$$

and

$$II = g(x) \cdot q(\text{load}) + 0.5[1 - q(\text{load})].$$

UD and II performance are now no longer simple functions of the underlying learning mechanism, but they also depend on the proportion of learners, itself a function of load. This defines the two-dimensional structure shown schematically in Figure 10 and will, under most circumstances, produce a bidimensional state-trace plot similar to that found by Z&M. In essence, the varying proportions of learners introduce a confound that makes it look as if multiple processes are present when they are not.

In order to see if the above analysis explains the apparent bidimensionality of the Z&M data, we applied the same accuracy criterion to their data and reanalyzed it. Excluding participants on the basis of the accuracy criterion ( $>65\%$  in the final block) reduces the  $N$  from 142 to 93 for Z&M’s Experiment 1. The proportions of “learners” were .78 and .34 for the UDC and UDS groups, respectively, and .88 and .59 for the IIC and IIS groups, respectively (see Table 2). Figure 11 shows the data for only the “filtered” participants. The standout feature of these data is the ordinal position of the UDS group. From being the poorest performers in Figure 5A, this group now performs as well as the UDC group by the end of the trials. Indeed, the data are characterized by the *absence* of a load effect; participants perform more poorly overall on the II than on the UD task, but there is no differential impact of load. An ANOVA confirmed these impressions, showing a significant main effect of category structure [ $F(1,89) = 24.96, p < .001$ ] but no effect of load [ $F(1,89) = 0.669, p > .40$ ] and no interaction of these factors ( $F < 1$ ). Moreover, if one examines the data for the excluded nonlearners, there are essentially no differences between conditions (average accuracy across the five blocks ranges from .50 to .53). This strongly suggests that the excluded participants were responding randomly.

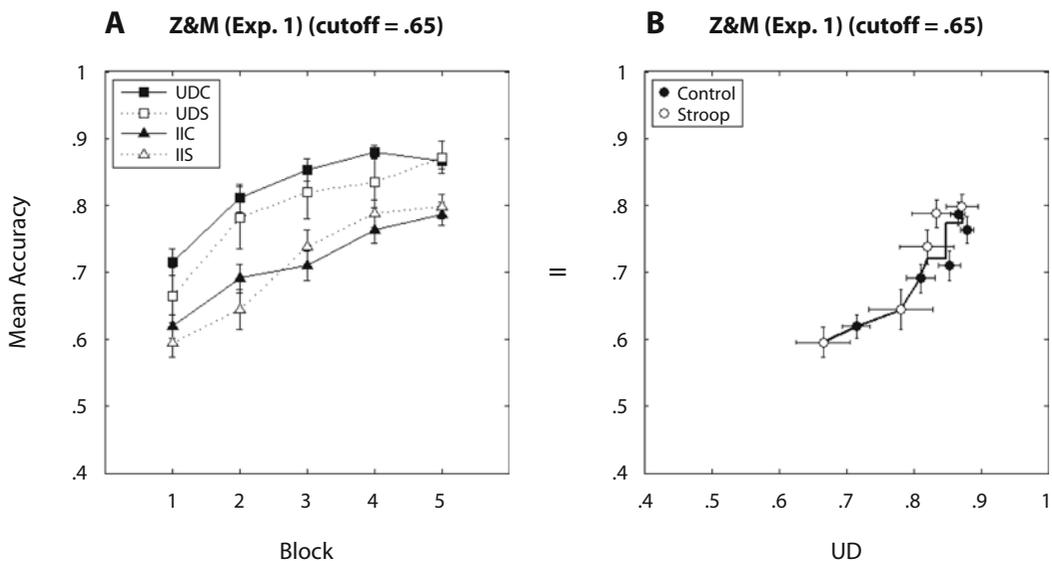


Figure 11. (A) Mean proportion correct for each group in Zeithamova and Maddox (2006), Experiment 1; only participants scoring .65 or above in the final block and .80 or above on the Stroop task are included ( $N = 93$ ). (B) State-trace plot of the data from panel A. The dark line shows the best-fitting monotonically increasing curve. UDC, unidimensional rule-based control; UDS, unidimensional rule-based Stroop; IIC, information integration control; IIS, information integration Stroop.

Figure 11B displays the corresponding state-trace plot for the filtered data from Z&M Experiment 1. The contrast with Figure 5B is stark. When only the filtered participants are considered, the apparent bidimensionality of the data disappears, with all points now occupying the same trajectory and lining up on a single monotonically increasing curve. The data are consistent with both the trace model ( $\Delta G^2 = 0.26, p = .899$ ) and the one-dimensional model ( $\Delta G^2 = 1.63, p = .555$ ), and the nonoverlap model can be strongly rejected ( $\Delta G^2 = 69.67, p < .001$ ). This is essentially the same pattern that was observed across our three experiments (cf. Figures 6, 7, and 8).

It is now clear that the apparent bidimensionality of the unfiltered data from Z&M Experiment 1 is due to the selective attrition of participants across the four groups. Since this attrition had a greater effect on the UDS condition, these data were effectively shifted to the left of their position in Figure 11B to produce the state-trace plot shown in Figure 5B. In contrast, in each of our three experiments, greater attrition occurred in the IIS condition, which had a much reduced effect on the state-trace plots (see Figures B1 to B3 in Appendix B for a comparison).

Our analysis of the effect of nonlearners leaves an obvious question unanswered: Why did we observe higher attrition in the II conditions under load in all three of our experiments, but Z&M found higher attrition in the UD condition in their experiment? It could be argued that the selective failure to learn by participants in the UDS group is itself a vindication of the COVIS model, which predicts this condition to be especially problematic. There are three difficulties with this view. First, it is unclear if any theory of category learning can be easily extended to account for the *failure* to learn. Second, even if selective attrition is a prediction of COVIS, it is one that we failed to replicate in three experiments. Third, it leaves unexplained why, even if the verbal system of some participants was disrupted by the additional cognitive load, the nonverbal system, operating in parallel, failed to master the task (see Zeithamova & Maddox, 2007, for speculation on this vexing issue). At the present time, we do not have sufficient information to satisfactorily explain the different patterns of attrition, which may well depend on uncontrolled factors such as individual differences in our participant populations, the vagaries of laboratory features or procedures, or simple chance.

In summary, the two results found by Z&M—a dissociation between the UD and II tasks, and an apparent bidimensionality of the state-trace plot—can both be attributed to a mixture of learning profiles due to the inclusion of nonlearners in the data analysis. This is a critical point and, as we have shown, may lead to spurious results, due to the fact that the proportion of nonlearners may vary across experimental conditions for reasons that have no bearing on models of category learning. When only learner data are considered, the qualitative pattern of results for all conditions across all the experiments—ours and Z&M's—is remarkably consistent. This consistency demonstrates that for the present combination of factors—cognitive load and number of trials—there is little or no evidence for more than a single underlying resource accounting for performance.

## GENERAL DISCUSSION

### State-Trace Analysis and Category Learning

Our aim has been to propose an alternative and justified procedure for examining claims of multiple processing systems underlying category learning. Our key contribution is a general one: highlighting why state-trace analysis offers the appropriate level of scrutiny for assessing claims made on the basis of functional and, very often, single dissociations. Our specific application of this approach to Z&M's data and experimental design has raised doubts about the need for a multiple-system interpretation in this particular context. Of course, this is just one manipulation and one data set, and the COVIS model is able to marshal a great deal of other experimental and neuroscience support (e.g., Ashby & Maddox, 2005; Filoteo et al., 2005; Maddox & Ashby, 2004). Therefore, our conclusions with respect to COVIS in particular, and the multiple-systems perspective in general, are similarly limited in scope. Nonetheless, we have highlighted a potential way in which researchers interested in this debate can draw stronger and more reliable conclusions from their data, whether those conclusions support single- or multiple-system interpretations.<sup>7</sup>

### The Impact of Filtering Data

An unexpected consequence of the present study was to reveal the impact of filtering data. The practice of removing participants who fail to learn has been relatively common and is well justified on the basis of increasing the sensitivity of the data analysis. Our application of state-trace analysis reveals an additional and critically important reason for excluding nonlearners. Variation in the proportion of nonlearners as a function of the variable of interest—in this case cognitive load—can have the effect of introducing a spurious additional variable to the system under consideration. Even if the effects of load and learning on performance on UD and II tasks is mediated by a single processing system, this can be masked by the inclusion of nonlearners. This result illustrates the general point that it is typically a trivial exercise to show that different tasks may depend upon different processes; all that is required is that the tasks be different enough. In the present series of experiments, contamination from different proportions of nonlearners across the experimental conditions was sufficient to induce an apparent bidimensionality in the data. An important contribution of state-trace analysis in the present case was to highlight this effect and to provide a framework through which it can be identified and understood. This effect underlines the necessity of isolating the system or systems of interest from other processes through careful experimental control and, in this case, by identifying and accommodating heterogeneity in the manner in which participants approached the category learning task.

### The Effect of Cognitive Load on Category Learning

Our reanalysis of Z&M and our three new experiments question the reliability and interpretation of a selective effect of increased cognitive load on learning UD struc-

tures. We chose to examine this manipulation because of the controversy already surrounding the interpretation of the load effect (Nosofsky & Kruschke, 2002; Stanton & Nosofsky, 2007; see also Lagnado et al., 2006; Newell et al., 2007; Tharp & Pickering, 2009). The overall picture that emerges from these studies is that even if a selective effect of cognitive load can be found, its interpretation is not straightforward, and certainly does not compel a multiple-system view. Our own results converge with this picture, but, more importantly, our approach demonstrates that experimental procedures that attempt to infer underlying processes from dissociations between UD and II tasks will never succeed, because dissociations are neither necessary nor sufficient for rejecting single-resource accounts.

## CONCLUSION

The goal of uncovering the processes underlying category learning is important for understanding how people perform such tasks. We have proposed that the basis for inferring the existence of multiple processing systems needs to be founded on a rigorous methodology that avoids misleading interpretations. In place of a reliance on simple dissociations that are open to criticism, we have proposed that state-trace analysis be used to determine the dimensionality of the underlying processing system. To support this strategy, we have developed a statistical procedure for identifying departures from monotonicity and have highlighted the need to isolate the critical learning system by removing contamination from processes external to the scope of the theory.

Our aim in applying state-trace analysis in the present context is, simply, to identify the circumstances in which categorization of UD and II structures depends upon a common underlying parameter representing a single source of information or aspect of mental processing and the circumstances in which it depends upon separate underlying parameters. We consistently found evidence that experience—operationalized as the number of training blocks—and processing load—operationalized using a variety of divided attention tasks—affect categorization performance of UD and II structures through the mediation of a single intervening parameter or resource. Any viable model of categorization must be able to account for this fact.

## AUTHOR NOTE

The support of the Australian Research Council (Grant DP: 0877510 awarded to the three authors) is gratefully acknowledged. We thank E.-J. Wagenmakers for his statistical advice, and Carissa Bonner, Emily Adcock, Katie Simmons, Rachel Stephens, and Anastasia Ejova for assistance with data collection. Correspondence concerning this article should be addressed to B. R. Newell, School of Psychology, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: ben.newell@unsw.edu.au).

## REFERENCES

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.

- ASHBY, F. G., & ELL, S. W. (2002). Single versus multiple systems of category learning: Reply to Nosofsky and Kruschke (2002). *Psychonomic Bulletin & Review*, **9**, 175-180.
- ASHBY, F. G., & LEE, W. W. (1991). Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General*, **120**, 150-172.
- ASHBY, F. G., & MADDOX, W. T. (2005). Human category learning. *Annual Review of Psychology*, **56**, 149-178.
- BADDELEY, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, **4**, 829-839.
- BAMBER, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, **19**, 137-181.
- BRAINARD, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, **10**, 433-436.
- DECARO, M. S., CARLSON, K. D., THOMAS, R. D., & BEILOCK, S. L. (2009). When and how less is more: Reply to Tharp and Pickering. *Cognition*, **111**, 397-403.
- DECARO, M. S., THOMAS, R. D., & BEILOCK, S. L. (2008). Individual differences in category learning: Sometimes less working memory capacity is better than more. *Cognition*, **107**, 284-294.
- DUNN, J. C. (2003). The elusive dissociation. *Cortex*, **39**, 177-179.
- DUNN, J. C. (2008). The dimensionality of the remember-know task: A state-trace analysis. *Psychological Review*, **115**, 426-446.
- DUNN, J. C., & JAMES, R. N. (2003). Signed difference analysis: Theory and application. *Journal of Mathematical Psychology*, **47**, 389-416.
- DUNN, J. C., & KIRSNER, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, **95**, 91-101.
- EFRON, B., & TIBSHIRANI, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
- FILOTEU, J. V., MADDOX, W. T., SIMMONS, A. N., ING, A. D., CAGIGAS, X. E., MATTHEWS, S., ET AL. (2005). Cortical and subcortical brain regions involved in rule-based category learning. *NeuroReport*, **16**, 111-115.
- FOERDE, K., POLDRACK, R. A., & KNOWLTON, B. J. (2007). Secondary task effects on classification learning. *Memory & Cognition*, **35**, 864-874.
- HARLEY, E. M., DILLON, A. M., & LOFTUS, G. R. (2004). Why is it difficult to see in the fog? How stimulus contrast affects visual perception and visual memory. *Psychonomic Bulletin & Review*, **11**, 197-231.
- HEATHCOTE, A., BROWN, S. D., & PRINCE, M. (2009). *The design and analysis of state-trace experiments*. Manuscript submitted for publication.
- HENSON, R. (2006). Forward inference using functional neuroimaging: Dissociations versus associations. *Trends in Cognitive Sciences*, **10**, 64-69.
- JAMESON, T. L., HINSON, J. M., & WHITNEY, P. (2004). Components of working memory and somatic markers in decision making. *Psychonomic Bulletin & Review*, **11**, 515-520.
- JANG, Y., & NELSON, T. O. (2005). How many dimensions underlie judgments of learning and recall? Evidence from state-trace methodology. *Journal of Experimental Psychology: General*, **134**, 308-326.
- KRUSCHKE, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22-44.
- LAGNADO, D. A., NEWELL, B. R., KAHAN, S., & SHANKS, D. R. (2006). Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General*, **135**, 162-183.
- LOFTUS, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, **6**, 312-319.
- LOFTUS, G. R., OBERG, M. A., & DILLON, A. M. (2004). Linear theory, dimensional theory, and the face-inversion effect. *Psychological Review*, **111**, 835-865.
- LOVE, B. C., & GURECKIS, T. M. (2007). Models in search of a brain. *Cognitive, Affective, & Behavioral Neuroscience*, **7**, 90-108.
- MADDOX, W. T., & ASHBY, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes*, **66**, 309-332.
- MADDOX, W. T., & ING, A. D. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 100-107.

- MCLACHLAN, G., & PEEL, D. (2000). *Finite mixture models*. New York: Wiley.
- MEDIN, D. L., & SCHWANENFLUGEL, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning & Memory*, *7*, 355-368.
- MURPHY, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.
- NEWELL, B. R., & DUNN, J. C. (2008). Dimensions in data: Testing psychological models using state-trace analysis. *Trends in Cognitive Sciences*, *12*, 285-290.
- NEWELL, B. R., LAGNADO, D. A., & SHANKS, D. R. (2007). Challenging the role of implicit processes in probabilistic category learning. *Psychonomic Bulletin & Review*, *14*, 505-511.
- NOMURA, E. M., & REBER, P. J. (2008). A review of medial temporal lobe and caudate contributions to visual category learning. *Neuroscience & Biobehavioral Reviews*, *32*, 279-291.
- NORMAN, D. A., & BOBROW, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, *7*, 44-64.
- NOSOFSKY, R. M., & JOHANSEN, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, *7*, 375-402.
- 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.
- NOSOFSKY, R. M., & ZAKI, S. R. (1998). Dissociations between categorization and recognition in amnesics and normal individuals: An exemplar-based interpretation. *Psychological Science*, *9*, 247-255.
- PRICE, A. L., FILOTEO, J. V., & MADDOX, W. T. (2009). Rule-based category learning in patients with Parkinson's disease. *Neuropsychologia*, *47*, 1213-1226.
- SEGER, C. A., & CINCOTTA, C. M. (2005). The roles of the caudate nucleus in human classification learning. *Journal of Neuroscience*, *25*, 2941-2951.
- SEGER, C. A., & CINCOTTA, C. M. (2006). Dynamics of frontal, striatal, and hippocampal systems during rule learning. *Cerebral Cortex*, *16*, 1546-1555.
- SHALLICE, T. (1988). *From neuropsychology to mental structure*. New York: Cambridge University Press.
- SHERRY, D. F., & SCHACTER, D. L. (1987). The evolution of multiple memory systems. *Psychological Review*, *94*, 439-454.
- STANTON, R. D., & NOSOFSKY, R. M. (2007). Feedback interference and dissociations of classification: Evidence against the multiple-learning-systems hypothesis. *Memory & Cognition*, *35*, 1747-1758.
- 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.
- WAGENMAKERS, E.-J., RATCLIFF, R., GOMEZ, P., & IVERSON, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28-50.
- 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.
- ZEITHAMOVA, D., & MADDOX, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, *34*, 387-398.
- ZEITHAMOVA, D., & MADDOX, W. T. (2007). The role of visuospatial and verbal working memory in perceptual category learning. *Memory & Cognition*, *35*, 1380-1398.

## NOTES

1. We are aware that this is only one of many tens of different manipulations that show dissociations between performance on II and UD tasks and acknowledge that our conclusions with regard to the COVIS research program are therefore limited in scope.

2. We are indebted to Dagmar Zeithamova for providing us with the raw data from Zeithamova and Maddox (2006), Experiment 1.

3. Our approach complements the Bayes factor approach to the same issue recently proposed by Heathcote, Brown, and Prince (2009).

4. The data for all participants scoring above .80 in the Stroop task are displayed in Appendix B, Figure B1A. The principal difference in the qualitative pattern is a reversal of the IIC and UDS conditions, with the former slightly outperforming the latter. This reversal had little impact on the statistical analyses: All reported main effects and interactions remained significant, with the addition of significant interaction between block and load reflecting steeper learning curves for the control over the Stroop conditions. Analysis of the state-trace plot of these data (Figure B1B) reveals that neither the trace model nor the one-dimensional model can be rejected ( $\Delta G^2 = 0.11$  and  $0.17$ , respectively), but the non-overlap model clearly fails ( $\Delta G^2 = 15.89$ ,  $p = .001$ ).

5. The data for all participants scoring above .80 in the Stroop task are displayed in Appendix B, Figure B2A. The qualitative pattern is similar to the filtered data. In the statistical analyses the interaction between category structure and load, which was marginally significant in the filtered data, was not significant. There is an additional significant interaction between block and load reflecting steeper learning curves for the control over the Stroop conditions. Analysis of the state-trace plot of these data (Figure B2B) reveals that neither the trace model nor the one-dimensional model can be rejected ( $\Delta G^2 = 0$  and  $0.06$ , respectively), and that the nonoverlap model clearly fails ( $\Delta G^2 = 30.11$ ,  $p < .001$ ).

6. The data for all participants scoring above .80 in the digit recall task are displayed in Appendix B, Figure B3A. The principal difference in these data is a reversal in the position of the IIC and UD-DR conditions. This reversal led to the effect of load being significant in these data [ $F(1,124) = 12.43$ ,  $p = .001$ ]. Analysis of the state-trace plot of these data (Figure B3B) reveals that neither the trace model nor the one-dimensional model can be rejected (both  $\Delta G^2 = 0$ ), and that the nonoverlap model clearly fails ( $\Delta G^2 = 21.15$ ,  $p = .001$ ).

7. In a recent investigation of the effects of delayed feedback on II and UD learning, we have found evidence consistent with more than one process (i.e., a 2-D structure). Thus, our failure to find bidimensionality in the present series of experiments is not due to the method per se.

## APPENDIX A

### Statistical Analysis of State-Trace Plots Using MLE/NHST

In this section, we describe our approach to the statistical analysis of state-trace plots based on maximum likelihood estimation (MLE) and null hypothesis statistical testing (NHST). A complementary approach based on Bayes factors has recently been described by Heathcote, Brown, and Prince (2009). In comparison with Heathcote et al.'s method, the MLE/NHST method has the advantage that it can be readily applied to the present data sets and does not require specification of prior probability distributions for each model. A disadvantage of our method is that the null hypothesis can only be rejected; in contrast to the Bayes factor approach, evidence for the null (e.g., for a unidimensional model) cannot be directly compared with evidence for alternative hypotheses (e.g., for a bidimensional model).

The MLE/NHST method consists of two parts: model fitting using constrained MLE and hypothesis testing in which the  $p$  values for each model fit are estimated using a Monte Carlo procedure based on the data-informed parametric bootstrap crossfitting method (PBCM) proposed by Wagenmakers, Ratcliff, Gomez, and Iverson (2004).

## APPENDIX A (Continued)

**Maximum Likelihood Estimation**

Within each condition, we assume that the set of observations (mean proportion correct for each participant) are independent and normally distributed with a common variance estimated by the weighted average of within-condition variance. Let  $x_j$  and  $y_j$  be the observed means for the  $j$ th condition in the UD and II tasks, respectively, and, for a given model, let  $\hat{x}_j$  and  $\hat{y}_j$  be the predicted means for the corresponding conditions. The fit of this model ( $G^2$ ) to the data is then given by the following equation:

$$G^2 = \sum_{j=1}^K \left( \frac{n_{xj}(x_j - \hat{x}_j)^2}{s_x^2} + \frac{n_{yj}(y_j - \hat{y}_j)^2}{s_y^2} \right),$$

where  $K$  is the total number of conditions (10 in the present case),  $n_{xj}$  and  $n_{yj}$  are the number of participants in the  $j$ th condition of the UD and II groups, respectively, and  $s_x^2$  and  $s_y^2$  are the corresponding pooled variance estimates.

We fit three models to the data: the *trace model*, the *one-dimensional model*, and the *nonoverlap model*. These models can be defined by the different constraints they impose on the condition means,  $\hat{x}_j$  and  $\hat{y}_j$ , which we represent as constraints on the signs of pairwise differences. Let  $u_{jk} = \hat{x}_j - \hat{x}_k$  be the difference between the means of conditions  $j$  and  $k$  for the UD task and let  $v_{jk} = \hat{y}_j - \hat{y}_k$  be the corresponding difference for the II task. The trace model specifies an ordering of conditions on both tasks. Specifically, within the control and Stroop conditions, it requires that performance be a nondecreasing function of blocks for both tasks. Each of these orderings can be defined in terms of the *signs* of the differences between the relevant conditions—that is,  $\text{sign}(u_{jk}) = \text{sign}(v_{jk}) = 1$ , for appropriate  $j$  and  $k$ . The one-dimensional model specifies an additional correspondence between *all* the pairwise differences on the two tasks—that is,  $\text{sign}(u_{jk}) = \text{sign}(v_{jk})$ , for all  $j$  and  $k$ . Finally, the nonoverlap model can be thought of as a one-dimensional model nested within a specific instance of the trace model, in which there is an additional order constraint between all the Stroop conditions and all the control conditions—that is,  $\text{sign}(u_{jk}) = \text{sign}(v_{jk}) = 1$  for all  $j \in \text{control}$  and  $k \in \text{Stroop}$ .

The different models were fit using the MATLAB FMINCON procedure. The starting point for each search was a vector of the observed means, although checks were implemented to avoid local minima.

**Null Hypothesis Statistical Testing**

When comparing nested models, the  $G^2$  statistic is distributed as chi-square with  $N - k$  degrees of freedom, where  $N$  is the number of data points and  $k$  is the number of model parameters. This does not apply in the present case, since the models differ in order constraints rather than number of parameters and, in addition, it is not possible to specify the number of such parameters a priori. For this reason, we tested the obtained  $G^2$  values against an empirical distribution derived for each experiment. We recently used (Newell & Dunn, 2008) a similar approach based on a procedure outlined by McLachlan and Peel (2000). However, this approach does not account for uncertainty in the parameter estimation. For this reason, we adopted a procedure similar to the PBCM method described by Wagenmakers, Ratcliff, Gomez, and Iverson (2004). This method was developed to compare two models of differing complexity. In our application, since the two models in question are nested within each other, we know that one is more complex than the other. Our approach is as follows. First, the data are resampled  $M$  times using the parametric bootstrap (Efron & Tibshirani, 1993). Second, the trace model is fit to each resampled data set and the best-fitting model parameters used to generate new data that are consistent with the trace model and also take initial parameter uncertainty into account. The trace model is then refit to these data, and the difference between the obtained  $G^2$  value and that of the unrestricted model in which it is nested is stored. This is a notional comparison only, since the  $G^2$  of the unconstrained model is necessarily zero. This process is then repeated for both the one-dimensional and nonoverlap models. In the former case, the one-dimensional model is fit to the resampled data and a data set generated from the best-fitting parameter values. The one-dimensional model and the trace model in which it is nested are then fit to these data and the difference in their  $G^2$  values stored. For the nonoverlap model, this model is fit to the resampled data and a nonoverlap data set generated from the resulting best-fitting parameter values. The nonoverlap model and the one-dimensional model in which it is nested are then fit to these data and the difference in  $G^2$  values stored.

For each hypothesis—trace model versus unconstrained model, one-dimensional model versus trace model, and nonoverlap model versus one-dimensional model—the procedure results in  $M$  values representing a random sample of the empirical distribution of the corresponding differences in  $G^2$ . In our analyses, we set  $M$  equal to 1,000. We defined the empirical  $p$  value of a difference in  $G^2$  obtained from the observed data as the proportion of values in the corresponding empirical distribution that are greater than or equal to this quantity.

**APPENDIX B**  
**Unfiltered Data Sets**

Figures B1–B3 display unfiltered data from Experiments 1–3, respectively, and include all participants scoring .80 or above correct on the concurrent memory task. In each figure, panel A shows the mean proportion correct across blocks on the category task for the four groups, and panel B shows the state-trace plot of the data in each respective panel A. See notes 4, 5, and 6 for discussion of these data.

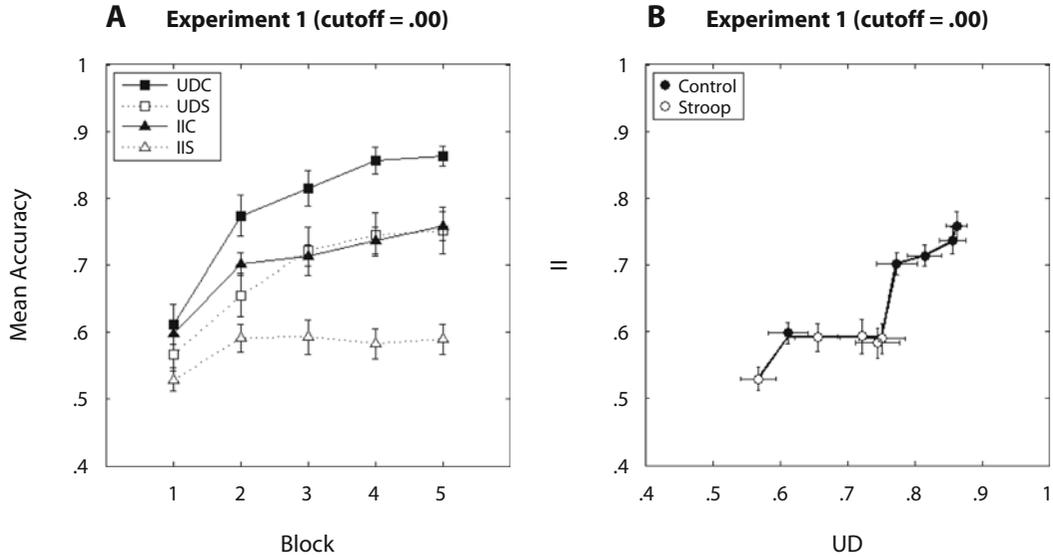


Figure B1. (A) Mean proportion correct for all conditions from Experiment 1; only participants scoring .80 or above on the Stroop task are included ( $N = 116$ ). The control groups are denoted with solid lines and filled marks, the dual Stroop task conditions with broken lines and unfilled marks. Error bars are standard errors of the mean. (B) State-trace plot of the data from panel A. The dark line shows the best-fitting monotonically increasing curve. UDC, unidimensional rule-based control; UDS, unidimensional rule-based Stroop; IIC, information integration control; IIS, information integration Stroop.

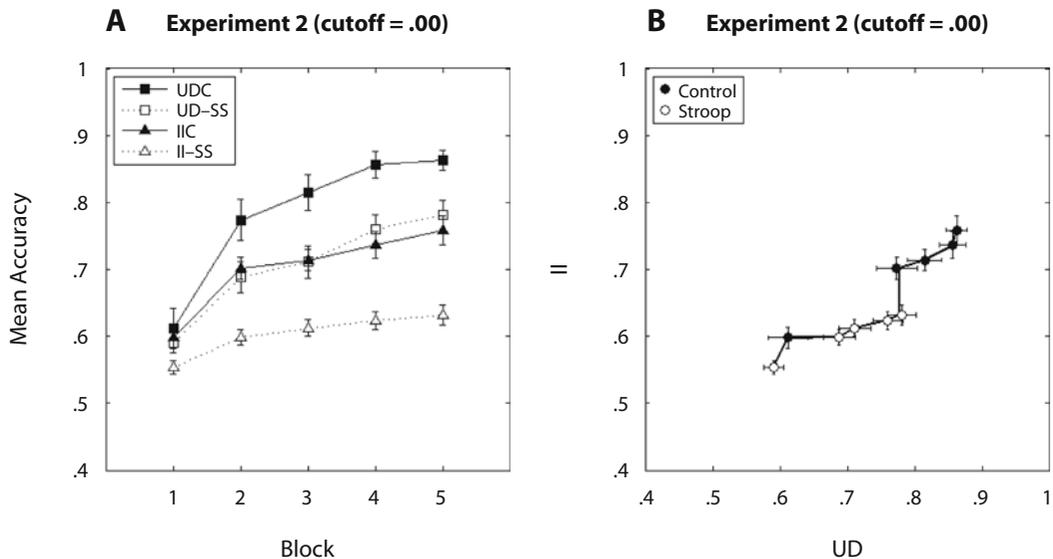
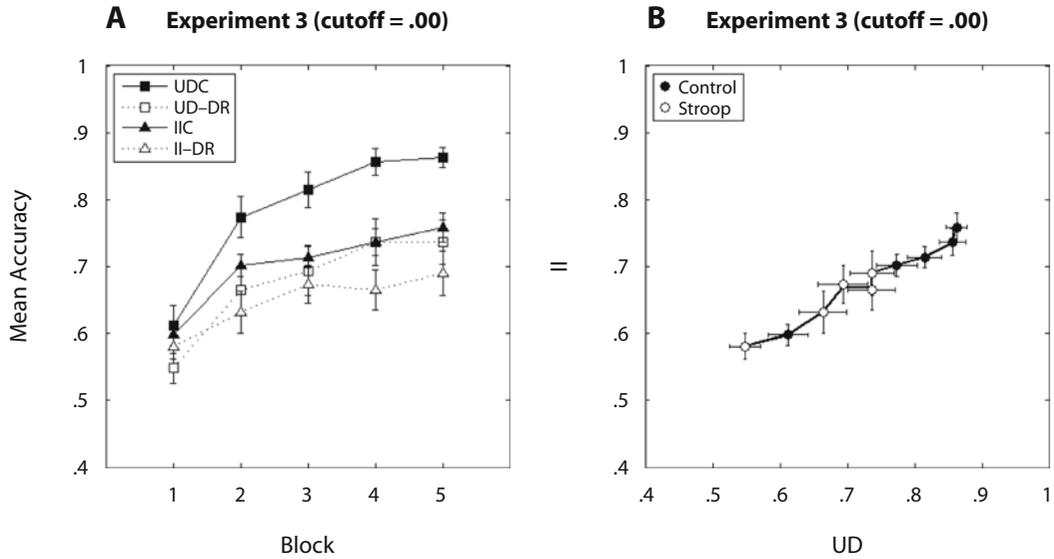


Figure B2. (A) Mean proportion correct for the simple-Stroop conditions of Experiment 2 and the control conditions from Experiment 1; only participants scoring .80 or above on the Stroop task are included ( $N = 184$ ). The control groups are denoted with solid lines and filled marks, the dual simple-Stroop task conditions with broken lines and unfilled marks. Error bars are standard errors of the mean. (B) State-trace plot of the data from panel A. The dark line shows the best-fitting monotonically increasing curve. UDC, unidimensional rule-based control; UD-SS, unidimensional rule-based simple Stroop; IIC, information integration control; II-SS, information integration simple Stroop.

APPENDIX B (Continued)



**Figure B3.** (A) Mean proportion correct for the dual-task digit recall conditions of Experiment 3 and the control conditions from Experiment 1; only participants scoring .80 or above on the digit recall task are included ( $N = 128$ ). The control groups are denoted with solid lines and filled marks, the dual digit recall task conditions with broken lines and unfilled marks. Error bars are standard errors of the mean. (B) State-trace plot of the data from panel A. The dark line shows the best-fitting monotonically increasing curve. UDC, unidimensional rule-based control; UD-DR, unidimensional rule-based digit recall; IIC, information integration control; II-DR, information integration digit recall.

(Manuscript received August 5, 2009;  
 revision accepted for publication January 5, 2010.)