Cognitive Science 29 (2005) 493–524

A Strategy-Based Interpretation of Stroop

Marsha C. Lovett

Department of Psychology, Carnegie Mellon University

Received 22 May 2004; received in revised form 18 January 2005; accepted 4 February 2005

Abstract

Most accounts of the Stroop effect (Stroop, 1935) emphasize its negative aspect, namely, that in particular situations, processing of an irrelevant stimulus dimension interferes with participants' performance of the instructed task. In contrast, this paper emphasizes the fact that, even with that interference, participants actually can (and usually do) exert enough control to perform the instructed task. An Adaptive Control of Thought–Rational (ACT–R) model of the Stroop task interprets this as a kind of learned strategic control. Specifically, the concept of utility is applied to the two processes that compete in the Stroop task, and a utility-learning mechanism serves to update the corresponding utility values according to experience and hence influence the competition. This model both accounts for various extant Stroop results and makes novel predictions about when people can reduce their susceptibility to Stroop interference. These predictions are tested in three experiments that involve a double-response variant of the Stroop task.

Keywords: Stroop; Strategy choice; Utility learning; Hybrid modeling; ACT–R; Computational modeling

1. Introduction

The Stroop effect (Stroop, 1935) is a long-studied, yet still intriguing, phenomenon in cognitive psychology. In its most general form, the Stroop effect occurs when two competing processes are relevant to the task at hand, but only one of these processes should govern the participants' response. The standard Stroop task requires naming the ink color of a word that, in some cases, spells a color (e.g., the word *red* printed in blue ink). The robust result is that performance varies as a function of the congruency between the ink color and the word. When the word spells a color that conflicts with the ink color, latencies and error rates increase relative to nonword and non-color-word controls, an effect known as Stroop *interference.* Conversely, Stroop *facilitation* occurs when the word spells a color that matches the ink color, with laten-

Requests for reprints should be sent to Marsha C. Lovett, Department of Psychology, Carnegie Mellon University, 124 Cyert Hall, Pittsburgh, PA 15213–3890. E-mail: lovett@cmu.edu

cies and error rates decreasing relative to controls. The most important aspect of Stroop phenomena is that these effects disappear under word-reading instructions.

Theories to explain Stroop effects abound. Each tries to explain people's *lack* of cognitive control, their inability to ignore the word when responding to the color. One explanation is that, through a lifetime of practice, reading has become so automatized that it impacts processing, even under color-naming instructions (Cohen, Dunbar, & McClelland, 1990). Other theories include response compatibility (Dalrymple-Alford & Azkoul, 1972), differential translation requirements (Virzi & Egeth, 1985), and speed of processing (Schooler, Neumann, Caplan, & Roberts, 1997). Evaluating these explanations often involves identifying and testing qualitative predictions they make. For example, a pure horse-race model has been discounted because results failed to support its prediction of a reversed Stroop effect when the ink color sufficiently precedes the word (Glaser & Glaser, 1982). Similar manipulations of stimulus onset asynchrony have been used to argue for modifications to a pure translation model of the Stroop effect (Sugg, & McDonald, 1994).

More recently, computational modeling has been especially productive in refining our understanding of the Stroop effect. First, quantitative predictions of performance, such as the relative size of particular Stroop effects, can more sharply distinguish among theories, especially in cases where qualitative differences between theories tend not to arise. Second, explanations of the Stroop effect often invoke theoretical constructs that require careful operationalization, which computational models by their very nature produce. Third, in the context of a cognitive architecture (cf. Newell, 1990)—where a set of fixed computational mechanisms applies *across* models—results from Stroop studies can constrain basic processes that are exercised in a broad range of other tasks.

1.1. Many effects, many models

The basic Stroop experiment has been varied in numerous ways, producing a diverse set of empirical effects that offer useful constraints for models of Stroop. In his 1991 review, MacLeod identified a list of Stroop-related findings that "must be captured by any successful theory [or computational model] of the Stroop effect" (p. 163). Table 1 presents a subset of that list, highlighting particularly influential results for testing computational accounts of the Stroop effect.

Given such a powerful set of modeling constraints, it has been a challenge to develop a model that meets all of them. Nevertheless, there have been many successful models fit to interesting subsets of the results (e.g., Altmann & Davidson, 2001; Botvinick, Braver, Barch, Carter, & Cohen, 2001; Cohen et al., 1990; Cohen & Huston, 1994; Phaf, van der Heijden, & Hudson, 1990; Roelofs, 2000, 2003; Roelofs & Hagoort, 2002). Three examples illustrate the range of computational approaches that have been taken to account for Stroop phenomena—two connectionist models (Cohen et al., 1990; Phaf et al., 1990) and a production-system model (Roelofs, 2000).

The Cohen et al. (1990) model and its extensions (e.g., Cohen & Huston, 1994; Botvinick et al., 2001) represent the two competing processes as separate pathways in a connectionist network. The competition between pathways is managed by task-control nodes that "gate" information processing and create a bias toward the instructed task. The central concept in this

Table 1
Empirical results that constrain models of the Stroop effect

Description of the Effect	Representative Citation(s)
The basic effect: The basic effect is robust to methodological variations, including list versus single-trial presentation, and task variants, such as the picture-word task. (1)	Dalrymple-Alford and Budayr, 1966
Semantic gradient: For noncolor words, the size of the Stroop effect increases with the strength of the semantic association to a color concept. (3)	Dalrymple-Alford, 1972
Facilitation: Facilitation can occur on congruent trials, but the size of this effect is smaller than interference and depends on the choice of neutral trials. (5)	various
Proportion of trial types: The proportion of trials of different types (conflict, congruent, and neutral) impacts the size of the Stroop effect. (7)	Tzelgov , Henik, and Berger, 1992
SOA: The maximal Stroop effect occurs when the color and word components of the stimulus are presented within 100 msec of each other. When the color precedes the word, a reverse Stroop effect (i.e., interference in the word-reading task) is *not* found. (10 & 11)	Glaser and Glaser, 1982
Degree of practice: The degree of practice at processing each of the two stimulus dimensions influences which task will interfere with the other and to what degree. (12)	MacLeod and Dunbar, 1988
Response modality: The modality of response matters with larger Stroop effects for oral rather than manual responses; also response compatibility impacts the size of Stroop effects. (13)	various

Note. In parentheses after each effect's description is the corresponding effect number from MacLeod's (1991) list of 18 major results. SOA = Stimulus-Onset Asynchrony.

model, *graded automaticity,* posits that greater practice at a task produces greater automaticity. The more automatized a task is relative to its competitor, the more its processing can proceed without additional input. Hence, word reading interferes with color naming even when the instructed task is color naming.

The Phaf et al. (1990) model was built as a Stroop extension to a connectionist model of visual attention. This model's architecture differs qualitatively between word reading and color naming in that there are direct input–output connections for word reading but not for color naming. This word-reading shortcut reduces the interference from color naming on word reading but allows for interference in the other direction.

The Roelofs (2000, 2003; Roelofs & Hagoort, 2002) model of Stroop phenomena was built as an extension to the WEAVER++ model of word production (Levelt, Roelofs, & Meyer, 1999) and specifies several stages of processing for word reading. Like the Phaf et al. (1990) model, it establishes a word-reading advantage by requiring fewer processing steps for that task, enabling different-sized congruency effects between word reading and color naming. This model was elaborated by Altmann and Davidson (2001) to include several aspects of Adaptive Character of Thought–Rational (ACT–R)'s declarative memory system.

1.2. A place for strategy in the Stroop task

What is immediately striking about the Stroop effect is that word reading *interferes* with color naming. And yet, despite this interference, participants answer the vast majority of trials correctly. Error rates range from 2% to 10% when the task involves naming the color of the ink. This implies that there is a mechanism governing the choice between the two processing pathways that takes into account the instructed goal. In the problem-solving literature, this is called *strategy choice* (e.g., Lovett, 1998; Siegler, 1991, 1996). In the context of the standard Stroop task, the word *strategy* refers to the procedure for processing the word or the ink color. It is worth noting that the word *choice* here need not, and usually does not, imply a decision made with conscious awareness. Although previous work on the role of strategy in Stroop (e.g., Cheesman & Merikle, 1984; Logan & Zbrodoff, 1982; Logan, Zbrodoff, & Williamson, 1984; Neill, 1978), has focused on explicit strategic approaches, this article focuses on the implicit strategy choice between word reading and color naming.

The literature on strategy choice in problem solving reveals that it is not only how much a strategy has been practiced that impacts its processing but how effective and efficient it was when applied (e.g., Lovett & Anderson, 1996; Siegler, 1996). One way to quantify "effectiveness and efficiency" is through a strategy's *utility.* This is a function of the strategy's expected gain—its likelihood of leading to success multiplied by the value attributed to that success—minus the expected costs of getting there. With *utility learning,* each strategy's associated utility value is updated based on its success and cost of achieving success when it is applied. A strategy with greater utility relative to its competitors will be more likely to be chosen in future situations where it is relevant. This utility-learning view includes practice as a part of effective strategy choice but further specifies that, in order for a strategy to become a more prepotent response, its applications must have led to high-success–low-cost outcomes—that is, high-utility outcomes.

Both a utility-learning model and a practice-based model, such as Cohen et al. (1990), specify a general mechanism for learning by experience that applies to diverse tasks and phenomena. Both predict that Stroop effects can occur between competing processes beyond word reading and color naming. Both predict that giving more practice to one of the competing processes will lead to Stroop effects. In the practice-based model, this is a direct effect of relative strength: The more practiced process will be more automatized and hence more prepotent. In a utility-learning model, extra practice tends to confer a utility advantage (e.g., practice leads to speedup, which reduces cost), leading to the same result. But a utility-learning model makes qualitatively different predictions from a purely practice-based model when the utility of the competing processes differs while the amount of practice is held constant. In this case, a utility-learning model predicts a Stroop effect, whereas a practice-based model does not.

This article describes a strategy-based, utility-learning model of Stroop and shows how it accounts for the results in Table 1 as well as new empirical findings that bear out the prediction just mentioned. The following section provides a description of this model and explains how it accounts for various extant Stroop results. Next, the article describes an infrequently studied variant of the Stroop task, called *double-response Stroop,* where participants are asked to respond to *both* aspects of the stimuli. Double-response Stroop is particularly interesting from a strategy-choice perspective because participants' relative preference for color-naming or word-reading can be observed on a trial-to-trial basis, for example, by observing which of the two stimulus dimensions is reported first. This double-response task is used to test this strategy- and utility-based model in a series of three experiments.

2. A strategy-choice model of Stroop

Most models of Stroop specify "universal" processing for each trial type in that the same steps or cycles are engaged regardless of the system's state (cf. Siegler, 1996). Although there is stochasticity in these models' processing (e.g., different responses and response latencies can be produced), the order in which information is processed and the kind of information that is processed is the same for every trial. In contrast, modeling Stroop effects as the result of strategy learning and choice—what this article calls a *strategy-based* perspective on Stroop—leads to a variety of ways that information can be processed; for example, different trials could focus on different kinds of information or process the same information but in different sequences. The difference between strategy-based and universal processing is akin to the difference between "fixed effects" statistical models, where the effects have fixed but unknown values, and "random effects" statistical models, where the effects are considered to be drawn from a population. Just as there is another layer of noise in "random effects" statistical models, so there is another level of (systematic) variation in strategy-based computational models. This extra variation in strategy-based models leads to the prediction of qualitative (not just quantitative) variation between individuals and even between individual trials completed by the same person. There are many examples of individual differences in when and whether Stroop effects occur (e.g., Chrisman, 2001; Comalli, Wapner, & Werner, 1962; Kane & Engle, 2003; Schiller, 1966). As for systematic trial-to-trial differences within individual participants, neuroimaging data is one source of evidence (Kerns et al., 2004).

Figure 1 presents a diagram of the different processing paths that this strategy-based model takes. This diagram highlights three branching points. One occurs early and diverts processing to focus on either the word or color dimension of the stimulus. Then, along both of these paths, there is another branching point that involves either immediately responding to the processed dimension or first checking that the task-relevant dimension has been processed. In addition, although it is not highlighted in this diagram, there is yet another opportunity for processing differences after task checking on both paths: When task checking reveals that the processed dimension matches the instructions, the model generates a response directly; otherwise, the model continues processing the stimulus, now focusing on the alternate dimension. This model allows for six qualitatively distinct processing pathways. Note that four of these involve task checking, which is quite likely to produce the correct response but takes extra time. The other two (depicted at the left and right extremes in Figure 1) save time by not checking but may respond with the task-irrelevant dimension—an error on conflict trials.

Besides supporting these qualitative predictions about strategy variation in Stroop, this model makes quantitative predictions that depend on the specific computational mechanisms driving the execution and learning of the model's strategies. The model represents strategies as *production rules*—contingencies for action of the form "When <conditions> are true, then do

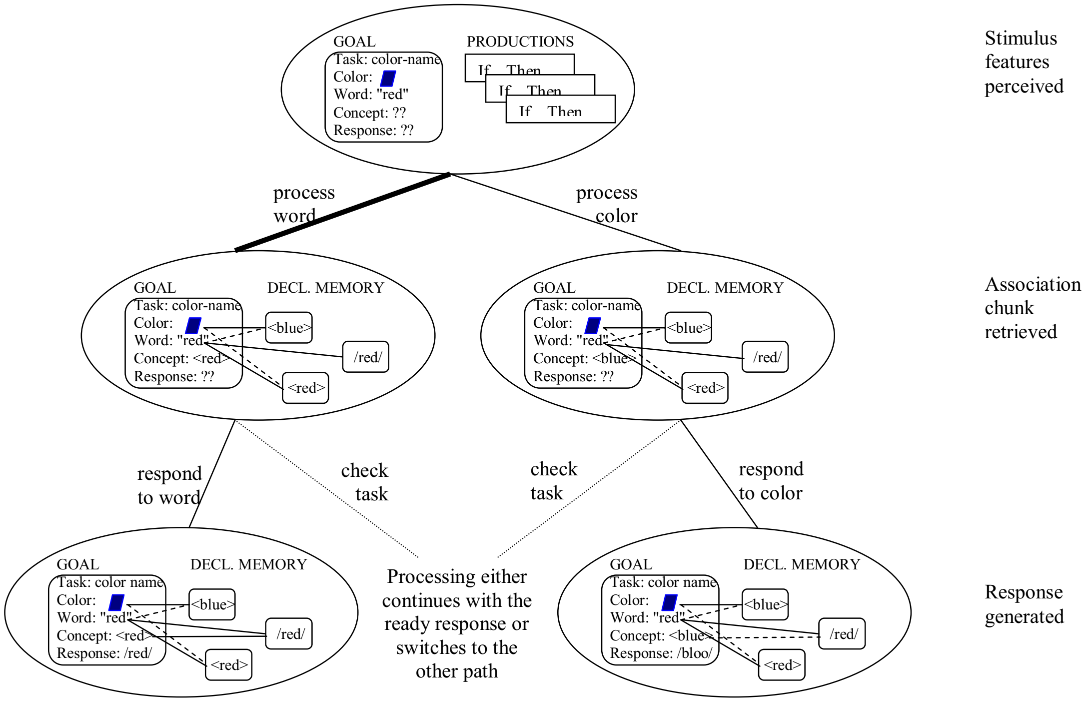

Fig. 1. Flow of control in this model for a color-naming trial. Lines between ovals represent production rules (thicker lines indicate greater likelihood of firing), and ovals represent parallel retrievals that modify information in the current goal. Along both the word-reading and color-naming paths, there is an optional check before responding (two lines labeled “check task”).

<actions>.” In this model, the color-naming production rule specifies as its condition that the task must be color naming. In contrast, the word-reading production rule has a very general condition; it will fire whenever the stimulus has wordlike features. So, when the task instructions specify color naming, both options match. However, the model only executes one production rule at a time, and that choice is governed by the utility values associated with the production rules: Generally speaking, the higher a production rule’s utility value relative to its competitors, the more likely it is to be executed. (In Figure 1, this is represented by different line thickness of the production rule lines.) The model’s computational mechanisms for using and updating utility values (discussed in the next subsection) enable specific quantitative predictions about how likely the model’s different processing paths are under various circumstances. For example, these mechanisms enable predictions about how likely it is that, say, a word response will be given without task checking when the task is color naming and there were 70% conflict trials.

When the model executes a particular production rule, it performs the actions associated with that production rule. Often this involves modifying the current goal based on information retrieved from memory (see ovals in Figure 1). For the process-word production rule, the goal is updated to reflect what is retrieved about the word dimension of the stimulus. That retrieval process involves a parallel competition among *word-association chunks,* so called because they link information from a word (the stimulus *red*) to its related concept (the concept of *redness,* or <red>), and possibly even to the associated motor program for reading that word (the verbal response /red/). For the process-color production rule, a similar action is performed, but there the retrieval involves a parallel retrieval of *color*-association chunks, each of which links a perceived ink color to the corresponding color concept. Whereas the competition among production rules is based on production rules' utility, the retrieval of chunks is based on a quantity called *activation,* which reflects how likely a chunk is to be needed, given its past use. The model's computational mechanisms for using and updating activation values (see next subsection) enable specific quantitative predictions about how likely different chunks are to be retrieved and how long each retrieval will take.

2.1. ACT–R implementation

The basic knowledge structures in this model are (a) production rules representing the various strategies participants employ in the Stroop task and (b) chunks representing the various facts and relations that these strategies require. The model is implemented in the cognitive architecture ACT–R (Anderson & Lebiere, 1998), which specifies a fixed set of mechanisms governing how such structures are used. In any ACT–R model, these same mechanisms are applied to derive quantitative predictions about learning and performance. In this Stroop model, three of ACT–R's mechanisms are particularly important: utility-based choice, utility learning, and activation-based retrieval. Each of these will be discussed in turn.

In each cycle of ACT–R's processing, a single production rule, from the set of those matching this situation, fires, and its actions are executed. ACT–R's utility-based choice mechanism specifies that the production rule with the highest utility, after some noise is added, is the one to fire. Specifically, the probability that production rule i fires depends on its utility (U_i), the utility of its competitors (U_j), and the distribution of the added noise (here, logistic noise with variance t):

$$P(i\ fires) = \frac{e^{-U_i/t}}{\sum_j e^{-U_j/t}}$$

In this model, this choice applies to each branching point in Figure 1. For the first choice, the process-word production rule has a higher utility than process-color, so the model is more likely to process the word dimension of the stimulus first. However, because the choice process is noisy, this will not always occur, producing qualitatively different trials, that is, those initiated by word reading versus color naming. Note that relative production-rule utility can be viewed as this model's operationalization of graded automaticity of processing: The higher one production rule's utility value is relative to another, the more likely it will dominate in the choice competition.

But where do these utility values come from? In ACT–R, utility of production rule i (U_i) is defined as $U_i = P_i * G - C_i$, where P_i is the estimated probability that production rule i leads to success, G is the value attributed to achieving success, and C_i is the estimated cost of achieving success (measured in units of time to complete this goal). Each time this goal is completed, ACT–R automatically updates the utility value for each of the production rules that fired as a part of that goal completion. This involves updating the two components of utility, *P* and *C,* for each relevant production rule *i* at that point in time *t:*

$$P_i\,(t) = [N_i\,(t-1)*P_i\,(t-1) + \text{CurrentSuccess}]/N_i\,(t)$$

$$C_i\,(t) = [N_i\,(t-1)*C_i\,(t-1) + \text{CurrentCost}_i]/N_i\,(t)$$

Here, $N_i\,(t-1)$ is the number of times production rule *i* had been fired up to time $t-1$, CurrentSuccess is an indicator variable for whether this goal was completed successfully, and CurrentCost_i is the time between production rule *i* firing and goal completion, that is, the time to achieve success for that production rule). Combining these utility-updating equations with ACT–R's choice mechanism implies that the more successful and less costly a production rule is in practice, the higher its utility value and hence the more likely it will fire. In general, these two mechanisms allow ACT–R models to learn to prefer strategies that are more effective and efficient. In this model of Stroop, if there were some kind of practice that would enable color naming to become more successful and less costly (or word reading to become less successful and more costly), the relative utility values would shift, thus increasing the likelihood that color naming fires, and reducing the prepotency of word reading. We test this prediction in Experiments 2 and 3.

The third ACT–R mechanism, activation-based retrieval, specifies the time to retrieve a chunk *i* as $F*\exp(-A_i)$, where A_i is the chunk's total activation and *F* is a latency-scaling parameter. But what is total activation? It is the sum of *base-level activation* and *source activation.* A chunk's base-level activation increases with practice and decreases with delay according to ACT–R's declarative learning mechanism (cf. Anderson & Lebiere, 1998). A chunk's source activation reflects its relevance to this goal and is computed via the network of links connecting chunks (positively and negatively) to each other. Source activation spreads from the goal along these links, giving extra activation boosts (or dips) to chunks that are relevant (or irrelevant) to this goal. Thus, total activation of chunk *i* is

$$A_i = B_i + \Sigma W_j S_{ji}$$

where B_i is the base-level activation of chunk *i,* W_j is the amount of activation spreading from component *j* of this goal, and S_{ji} is the $i \rightarrow j$ link strength.

In this model, base-level activation for word-association chunks is preset to be higher than that for color-association chunks (2 vs. 0), reflecting greater prior practice at retrieving word-related information. In addition, the model was given the following fixed link strengths: (a) +1.5 for each pair of chunks involving matching colors, (b) –1.5 for each pair of chunks involving mismatching colors, (c) +0.6 for each pair of chunks involving matching tasks, (d) –0.6 for each pair of chunks involving mismatching tasks, and (e) 0 for all other pairs. Because the stimulus is represented in the goal, source activation spreads from the two stimulus dimensions along these links to various chunks in memory. In conflict (congruent) trials, this reduces (increases) the target chunk's total activation because of negative (positive) link strengths just mentioned.

2.2. How does the model account for basic Stroop?

The model's production-rule choice mechanism favors word reading (even when the task is color naming) because the process-word production rule has high initial utility. This means that the word will likely be processed (at least to some degree) first. Because word-association chunks have high base-level activations, the time required for their retrieval will be short. Moreover, the source activation increase (decrease) from congruent (conflict) trials will have little effect on retrieval time, because the retrieval time function is nonlinear. In some cases, the model will simply respond at this point, making word-reading trials highly accurate and fast, with little effect of trial congruency.

If the task is color naming, however, responding at this point will produce an error for conflict trials and a success for congruent trials. Thus, the model predicts more errors on conflict trials for the color-naming task. If, instead of responding immediately, the check-task production rule fires, a second pass of processing will be initiated. The relevant color-association chunk will be retrieved with a latency determined by its total activation. For congruent trials, total activation is boosted from source activation spreading from the stimulus color, the matching word, the current task, and the previously retrieved, matching word-concept. For conflict trials, source activation contributions to total activation are mixed: There are additions from the stimulus color and current task but reductions from the mismatching word and word concept. Because color-association chunks have moderate base-level activations and because the latency function is nonlinear, source activation contributions have a significant impact on color associations' retrieval times. Thus, for color naming, the model responds faster for congruent trials than for conflict trials, producing Stroop facilitation and interference.

It is worth noting that this model is consistent with a recent view that Stroop facilitation reflects participants sometimes making word-based responses on congruent, color-naming trials (MacLeod & MacDonald, 2000). Related work highlights the fact that, empirically speaking, Stroop facilitation is not the exact inverse of Stroop interference (e.g., MacLeod, 1998). In this model, Stroop facilitation stems from two influences—source activation boosts to memory retrieval and the possibility of responding quickly with the irrelevant dimension—the second of which does not have its inverse in Stroop interference.

2.3. Basic model fits

Besides explaining the model's functioning in the standard case, it is important to demonstrate that the model fits key results in the Stroop literature. Lovett (2002) showed that a preliminary version of this model accounted for five experiments containing 92 data points. This fit was performed over all 92 data points by varying 12 parameters; it produced an R^2 of .95 and mean deviation of 33.4 msec. The 12 varied parameters included a separate latency factor F and intercept for each of the five experiments plus an extra free parameter in two cases where atypical stimuli were used. These five experiments correspond to rows 1, 2, 5, and 6 from Table 1 plus an additional study indicating magnified Stroop interference among schizophrenic patients.

Here, we present the model's account for one experiment from that set—Macleod and Dunbar (1988)—because it is most relevant to strategy learning. In this experiment, participants learned a shape-naming task where different shapes were assigned color-word names. Over the course of

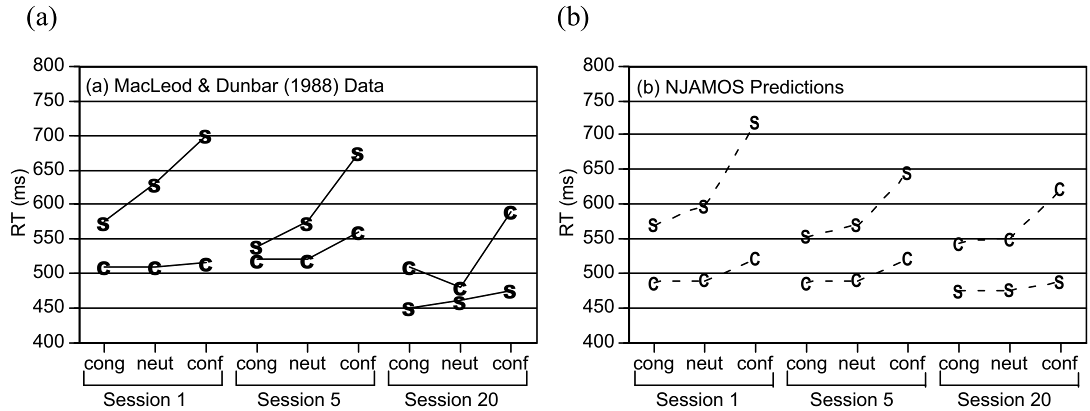

Fig. 2. Data from MacLeod and Dunbar (1988) (a) and this model's predictions (b). In each panel, RTs for shape naming (S) and color naming (C) are plotted separately across congruency and training manipulations.

20 days of practice at the shape-naming task, participants were tested intermittently in a Stroop task variant where shape naming and color naming compete (i.e., the shapes were presented in different ink colors to create congruent, neutral, or conflict trials under either shape-naming or color-naming instructions). From the data in Figure 2a, one can see that, at Day 1, color naming interfered with shape naming but not vice versa. At Day 20, these effects were reversed, and at Day 5 the results showed moderate interference in both directions.

Figure 2b presents our model's fit to these data. Note that this is a case where, besides the latency factor *F* and a general intercept, an additional parameter was used to represent the initial base-level activation for shape association. From this initial value, ACT–R's declarative chunk-learning mechanism was applied to update the base-level activations as the model received additional practice at shape naming during training. Note that increases in base-level activation make the shape chunks faster to be retrieved and less susceptible to trial-type effects. Also, over the course of training, the model updates the utilities of the shape-processing and color-processing production rules based on their effectiveness and efficiency in use. The change is quite dramatic for the shape-processing production rule because this is a completely novel task, and there is a lot of room for utility improvement. Specifically, the process-shape production rule's utility rises as it gets faster (because of the chunk learning mentioned earlier). This, in turn, makes the model more likely to process the shape first (because of utility-based choice). This strategy shift is another means by which the model speeds up in its shape naming across training; the model is no longer starting the shape naming after some initial color processing. Moreover, processing the shape first increases the size of the trial-type effect for color naming because the retrieved shape information spreads source activation to the relevant color-association chunk—a positive contribution on congruent trials and negative contribution on conflict trials.

Although MacLeod and Dunbar's (1988) experiment showed that practice at a nondominant task in a Stroop-like competition can reverse initial Stroop effects, practicing the standard Stroop task with equal proportions of the three trial types generally does *not* lead to a sizable reduction in the Stroop effect. This suggests that to produce noticeable change in participants' control of processing in the Stroop task, a more powerful impetus to learning must be established. This was accomplished via a new skill in MacLeod and Dunbar and has also been accomplished by increasing the proportions of conflict trials in various other studies (Cheesman & Merikle, 1984; Lindsay & Jacoby, 1994; Logan, 1980; Logan & Zbrodoff, 1979; Tzelgov, Henik, & Berger, 1992). In these studies, the general result is as follows: The more frequent the conflict trials, the smaller the size of the interference effect. Indeed, Experiment 3 from Stroop's original article gave participants repeated sessions of 100% conflict trials and produced both a striking reduction in the Stroop effect and even a reverse effect, that is, color naming interfered with word reading.

Although fits are not presented here, this model can account for such results nicely without further elaboration. The model simply updates the relative utility values of color naming and word reading through experience. The utility of color naming starts lower than word reading, but with a higher frequency of conflict trials, the utility value of color naming rises more. This leads to a higher probability of processing color information first under color-naming instructions and hence lowers the size of Stroop interference.

2.4. Logic of the experiments, task variant, and manipulations

Like MacLeod and Dunbar's (1988) shape-naming task and the manipulation of conflict trials' frequency, the experiments described in this article present participants with a more-difficult-than-usual Stroop situation, with the intention of impacting Stroop interference by producing a greater impact on the competing strategies' utility. Specifically, the following three experiments make use of the double-response Stroop task (Greenwald, 1972; Klein, 1964; Schweickert, 1978; Shimada & Nakajima, 1991) in which participants are shown standard Stroop trials but asked to respond to *both* stimulus dimensions. This task was chosen not only for its potential to produce larger utility changes but because it is particularly informative to a strategy-based view of Stroop, in that one can observe whether a participant responds to the word first or the color first and get a window onto potential strategy choices.

Past research on the double-response Stroop task has focused on the effects of response orders and response-modality mappings. Klein (1964) found that a word-first response order was easier than a color-first response order, but Shimada & Nakajima (1991) found no such difference. (Because both studies used 100% conflict trials, there is no assessment of the relative size of Stroop effects for the two response orders.) Greenwald (1972) and Schweickert (1978) found that a high-compatibility response-mapping (press button for color, say word) was easier than a low-compatibility mapping (press button for word, say color) and that this effect interacted with Stroop interference.

The double-response task experiments presented here touch on some of the same issues but focus on implications of our model of Stroop. In particular, the first experiment manipulates response-modality mapping (similarly to Greenwald, 1972, and Schweickert, 1978) but does so in a context where participants' response orders are not constrained. This enables a test of whether there is systematic strategic variation in response order, as a strategy-based view of Stroop would predict. The second and third experiments manipulate response order (similarly to Klein, 1964, and Shimada & Nakajima, 1991) but do so in a context where Stroop interference can be measured. This enables tests of specific predictions about change in Stroop interference that stem from the utility-based learning mechanism employed in this strategy-based model of Stroop.

3. Experiment 1

The goal of Experiment 1 was to explore what strategies participants naturally generate and exercise in the Stroop task. With the double-response task, one can measure, on a trial-by-trial basis, whether a participant responds to the word first or the color first. Moreover, if strategic differences are observed between individuals (or across trials within individuals), this experiment can test whether those differences are related to Stroop interference. If so, this would provide preliminary evidence that strategies do play a role in modulating Stroop interference, both directly (as measured in the double-response blocks) and indirectly (as measured in the standard Stroop blocks).

3.1. Methods

3.1.1. Participants

The participants in this experiment were 36 Carnegie Mellon University undergraduates who received course-related credit for participating. Participants were screened to confirm that they were right-handed, English speakers with full-color vision.

3.1.2. Design

The design of this experiment included one between-subject factor, namely whether the participants were instructed to respond to the color dimension manually and the word dimension verbally or vice versa. Seventeen participants were randomly assigned to the condition "Key in the color and say the word," and 19 participants were randomly assigned to the condition "Say the color and key in the word." As in the Greenwald (1972) and Schweickert (1978) studies of the double-response Stroop task, responding to the color manually and the word verbally can be considered the more natural or *high-compatibility* mapping, whereas responding to the color verbally and the word manually can be considered the less natural or *low-compatibility* mapping. Note that in *neither* condition were participants instructed as to the order in which they should make their responses in the double-response task. Instead, they were told (a) to respond as quickly as they could while maintaining accuracy and (b) to make each response as soon as they were able (i.e., to avoid "bundling" their responses).

The other factors included in this experiment involved the within-subject manipulations of task type and trial type. Specifically, there were four different tasks—color-only, word-only, standard Stroop, and double-response Stroop, each organized into blocks of 48 trials. The sequence of blocks was presented as three superblocks, each composed of the following: one-color-only block, one-word-only block, one standard Stroop block, and three double-response Stroop blocks. That is, participants completed 18 blocks (3 superblocks of 6 blocks each) at 48 trials each, for a total of 864 trials. Blocks were randomly ordered within each superblock for each participant. For the standard Stroop and double-response Stroop blocks, there were equal numbers of conflict and congruent trials.

3.1.3. Procedure

After a brief questionnaire to confirm eligibility and to collect demographic data, participants were given a brief overview of the study. Specific task instructions were presented on the computer screen using E-Prime software (Psychology Software Tools, Pittsburgh, Pennsylvania), and some of these were read aloud by the experimenter for emphasis. For example, participants were instructed differently, depending on their condition, as to how they should respond using the button box and by speaking into a voice-activated microphone. After these instructions were read and any questions answered, the experimental trials began. A task-specific instruction screen preceded each involving a new task.

3.1.4. Analysis

Reaction-time data were analyzed by computing the median response time for each participant over the relevant set of trials (e.g., a given block), for correct trials only. These medians were submitted to subsequent analyses, with aggregate results typically presented as means of these medians. Accuracy data were computed as the mean error rate for each participant over the appropriate set of trials. In all analyses, statistical significance was calculated based on an alpha level of 0.05. Note that for Stroop interference effects, the difference between conflict and congruent trials is used because this design did not include any neutral trials.

3.2. Results and discussion

The main findings of this experiment are (a) that participants *did* vary in their chosen order of response in the double-response task and (b) that these self-selected participant groups exhibited significantly different Stroop interference in the *standard* Stroop task. In this section we will describe these and other basic results in detail, organized by task type.

Figure 3 presents the reaction-time data for both response-mapping conditions and all four task types (with conflict and congruent trials presented separately for the two Stroop tasks). Corresponding accuracy data were very high—above 97% everywhere except for the double-response conflict trials in the say color-key word condition, where accuracy was 95%.

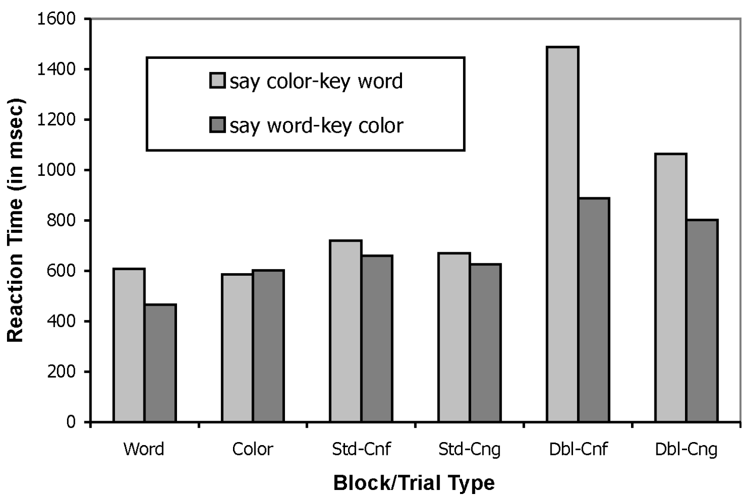

Fig. 3. Experiment 1 reaction-time data for both conditions and all four block types (conflict and congruent trials treated separately).

Thus, analyses of variance (ANOVAs) were conducted on the reaction-time data only. These analyses investigated, for each task, the effects of response-mapping condition, repetition (across three or nine blocks, depending on the task), trial type (for standard and double-response Stroop tasks only), and their interactions. The columns in Table 2 present the results of these analyses by task type, whereas the rows summarize whether each factor (or interaction) was significant across task types.

As the first row of Table 2 indicates, there was a difference in reaction time between the two conditions for all tasks except the color-only task. Not surprisingly, in the three tasks showing this effect, the high-compatibility condition responded significantly faster. The absence of this effect for the color-only tasks suggests that—regardless of response mapping—participants were consistently slow to respond to the ink color, even when it was the only stimulus dimension that they were either saying or keying.

Referring to the second row in Table 2, all four tasks showed a main effect of block repetition. For the color-only and double-response Stroop tasks, this main effect reflects a speedup across blocks. For the word-only and standard Stroop tasks, there was a speedup between Blocks 1 and 2 but then an increase in reaction times for Block 3. It is unclear why these reaction times would rise in Block 3 after a predictable decrease from Block 1 to Block 2, unless perhaps fatigue set in. It is also worth noting that block repetition did not show a significant interaction with condition for any of the tasks (see Row 3 of Table 2).

For the standard and double-response Stroop tasks, there are additional results involving trial type (conflict vs. congruent). Not surprisingly, the main effect of trial type was significant in both tasks, with conflict trials taking longer. This reflects the presence of Stroop interference

Table 2
ANOVA results for Experiment 1

	Task Type			
Factor	Word-Only	Color-Only	Standard Stroop	Double-Response
Response Mapping	$F(1, 34) = 63.9$**, $MSE = 8682$	$F(1, 34) = 0.49$ $MSE = 14152$	$F(1, 34) = 3.9$*, $MSE = 35900$	$F(1, 34) = 77.5$***, $MSE = 400400$
Block Repetition	$F(2, 68) = 19.4$**, $MSE = 1213$	$F(2, 68) = 3.9$*, $MSE = 1426$	$F(2, 68) = 12.0$**, $MSE = 4280$	$F(8, 272) = 60.7$***, $MSE = 15566$
RM × BR	$F(2, 68) = 1.7$, $MSE = 1213$	$F(2, 68) = 1.7$, $MSE = 1426$	$F(2, 68) = 0.9$, $MSE = 4280$	$F(8, 272) = 0.7$, $MSE = 15566$
Trial Type	n/a	n/a	$F(1, 34) = 45.4$***, $MSE = 2215$	$F(1, 34) = 91.3$***, $MSE = 117600$
TT × RM	n/a	n/a	$F(1, 34) = 1.8$, $MSE = 2215$	$F(1, 34) = 39.5$***, $MSE = 117600$
TT × BR	n/a	n/a	$F(2, 68) = 22.5$***, $MSE = 1794$	$F(8, 272) = 4.6$***, $MSE = 7000$
TT × RM × BR	n/a	n/a	$F(2, 68) = 5.4$**, $MSE = 1794$	$F(8, 272) = 2.4$*, $MSE = 7000$

Note. ANOVA = analysis of variance; *MSE* = mean square error; RM = Response Mapping; BR = Block Repetition; TT = Trial Type; n/a = not applicable.

*$p < .05$. **$p < .01$. ***$p < .001$.

in both tasks. It is interesting to note that, for the double-response Stroop task, there was also a significant interaction between trial type and condition. This interaction stems from the size of the Stroop interference effect being larger in the low-compatibility condition. This is not very surprising if we consider that the size of Stroop interference is simply “scaling” with the overall reaction time, which is so much larger for the low-compatibility participants in the double-response task. Also, this result is consistent with the findings of Greenwald (1972) and Schweickert (1978).

Other effects in the standard and double-response Stroop tasks involve the interaction between trial type and block repetition and the three-way interaction of Trial type × Block × Condition. The Trial type × Block interaction is significant for both Stroop tasks and likely reflects a general shrinking of the trial-type effect across blocks, which would be another example of Stroop interference scaling with overall reaction time. Finally, the three-way interaction is significant for both Stroop tasks. Although this high-order interaction is by its nature complex, the simplest interpretation is that the size of the trial-type effect (i.e., Stroop interference) scales with the overall reaction time, which is much greater for the low-compatibility condition and also much greater in the earlier blocks of the experiment.

Drawing firm conclusions about these Stroop interference effects, however, may not be warranted given the result foreshadowed earlier, namely, that Stroop interference effects were different for groups of participants who naturally exhibited different response orders in the double-response Stroop task. The next subsection describes this form of strategic variation, as observed in the double-response task, and presents a strategy-based analysis of Stroop interference effects.

3.2.1. Different strategies in the double-response Stroop task

The strategy analysis conducted on the double-response Stroop data involved assessing, for each double-response trial, whether the participant responded with the color dimension or the word dimension first. Recall that participants had free choice to respond word-first or color-first on a trial-by-trial basis. Figure 4 shows, for 6 selected participants in the high-compatibility condition, the proportion of word-first trials in each double-response block. There are several features of these data worth noting. First, the vast majority of the participants—31 out of 36 across both conditions (4 out of the 6 participants shown)—responded predominantly in a single order; that is, more than 90% of the time word-first or more than 90% of the time color-first. This is a striking result because there is no logical constraint of the task to choose either response order nor to *maintain* a previously chosen response order. Second, within the low-compatibility response-mapping condition (not depicted in Figure 4), the vast majority of participants (15 out of 17) were categorized as responding predominantly color-first. For these participants, color was reported verbally, so this result may reflect that the verbal response mode is either faster or more dominant. Third, within the high-compatibility condition, approximately equal numbers of participants were categorized as either responding predominantly color-first (7 out of 19), predominantly word-first (7 out of 19), or “mixed” (5 out of 19). Even among the 5 “mixed” participants, their choice of one response order or the other was fairly consistent *within each block.*

Given that the double-response Stroop data enable a categorization of participants according to their chosen response order, it is possible to reanalyze participants’ Stroop interference

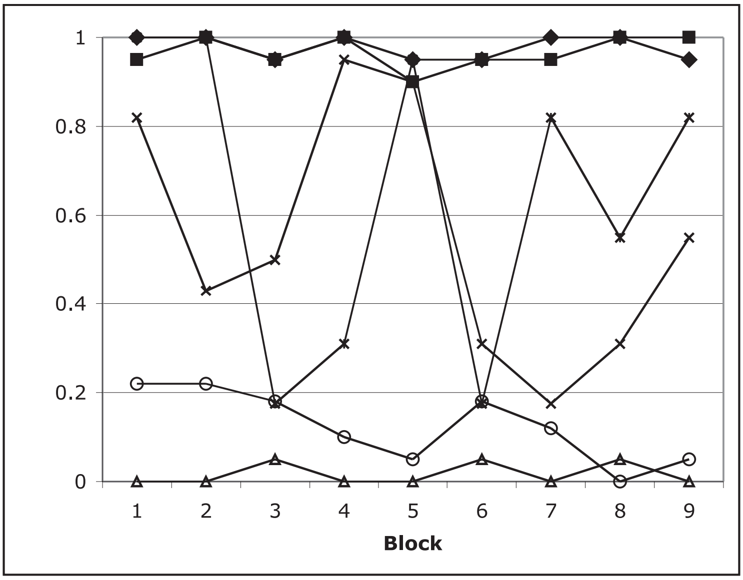

Fig. 4. Selected participants' response-order proportions from Experiment 1, presented by block, in the double-response task. Note that two of the participants predominantly responded word first (filled shapes), two predominantly responded color first (open shapes), and two showed mixed response orders.

effects separately for these groups. Interestingly, there is *no* reliable difference among the response-order groups in terms of the size of their *double-response* Stroop interference, $F(2, 16) = 2.4$, mean square error [MSE] = 1113. This may reflect some degree of adaptivity in participants' response-order choices: They use a response order that meshes well with their skills and hence, any differences in Stroop interference are attenuated.

What is more interesting is that there is a difference among the response-order groups in terms of their standard Stroop interference. Focusing on the participants in the high-compatibility condition (because there is greater strategic variability in this group), participants who choose to respond to the color first in the double-response blocks show the least Stroop interference in the standard blocks (less than 0 msec on average). Participants who chose to respond to the word first in the double-response task show the greatest Stroop interference in the standard Stroop blocks (44 msec), and participants in the mixed group are intermediate (26 msec). In other words, participants who choose a double-response strategy that involves reporting the color first may be shielding themselves from interference in the standard Stroop task. Indeed, this group's interference effect is significantly smaller than the other two groups', $F(1, 16) = 4.0$, $MSE = 4580$, an effect that will be further explored in Experiment 2.

It should be emphasized that participants are self-selecting into these response-order groups, that is, this is not an experimental manipulation. Therefore, these results cannot distinguish between the following two interpretations. One interpretation is that the double-response task acts as a sensitive diagnostic regarding the variability in people's natural approach to Stroop tasks, that is, the degree to which they emphasize color in their processing and responding. Under this interpretation, the difference in Stroop interference across response-order groups simply reflects individual variation that is a common cause of Stroop interference magnitude and response-order choice. Consistent with this interpretation is the fact that color-only reaction time correlates significantly with proportion word-first responses (among the high-compatibility condition where there is sufficient response-order variability), $r = .323$, $p < .05$.

The second interpretation is that the double-response task acts as an effective manipulation to change participants' relative emphasis on the two stimulus dimensions, which they can then transfer to the standard Stroop task, hence impacting Stroop interference. For example, by practicing color-first responding in the double-response task, one is becoming more likely to emphasize the ink color and thus will be able to perform the standard task with less interference.

Note that the model presented in this article is consistent with *both* interpretations. If participants are tending toward a particular response order for the double-response task, they will show the same bias in processing for the standard task, which will be reflected in their standard Stroop interference. The model would account for these related effects by positing individual differences in participants' a priori utility values for color naming and word reading. If, on the other hand, participants are more randomly choosing (and maintaining) a response order, the practice they get in the double-response task will impact their word-reading and color-naming production rule utilities in such a way they will be biased toward the same order of processing for the standard Stroop task. Regardless of which interpretation holds, the model specifically predicts the latter effect, namely that response order in the double-response task will impact standard Stroop interference. This prediction is tested in the next experiment.

4. Experiment 2

The goal of Experiment 2 is to use an experimental manipulation of participants' response order in the double-response Stroop task to test whether the relation between response order and size of the standard task Stroop effect, found in Experiment 1, is a causal one. That is, does practicing the double-response Stroop task in one order versus the other differentially impact the Stroop effect (as measured in the standard task)?

The current model predicts that the color-first order will significantly reduce the size of participants' Stroop effect and that the word-first order will moderately increase the size of participants' Stroop effect. This difference is predicted by this model because of the utility-based learning it employs. Note that a practice-based model would not predict a difference because either double-response order involves the same amount of practice. Under a utility-learning model, the assumption is that participants come into the experiment with a higher utility associated with word reading than color naming, thus tending toward a prepotent word-reading response. If participants bring to the experiment a bias toward this word-first strategy and yet are assigned to the color-first condition, word-reading will incur a very high cost as participants essentially engage a word-then-color-then-word sequence of processing, where the first attempt at word processing is something of a false start. That extra cost will negatively impact the word-reading production rule's utility over the course of double-response practice, making word reading less likely to fire first, and hence producing smaller Stroop interference in subsequent standard Stroop blocks. In contrast, for participants assigned to the word-first condition, the initial tendency to respond word-first will be rewarded, that is, for them, reading the word first is an immediate success. This may lead to an increase in the utility of word reading, but because it is already quite high relative to color naming, the model does not predict as large an increase in Stroop interference in this condition as it does a reduction in the other, color-first condition.

4.1. Methods

4.1.1. Participants

The participants in this experiment were 41 Carnegie Mellon University undergraduates who received course-related credit for participating. One additional participant was not included in the analysis for failure to follow instructions.

4.1.2. Design

The design of this experiment includes one between-subject factor called *order,* which specifies the order in which participants were instructed to respond to the two stimulus dimensions in the double-response Stroop blocks—either color then word or word then color. In this experiment, all participants used the same response modalities as the high-compatibility condition from Experiment 1: saying the word and keying in the color. Thus, color-then-word participants were always keying-then-saying their responses, and word-then-color participants were always saying then keying.

The within-subjects factors in this experiment correspond to the different blocks participants completed—all in a fixed order: word-only, word-only, color-only, color-only, standard Stroop, standard Stroop, double-response (eight contiguous blocks), standard Stroop, standard Stroop, color-only, word-only. This design enables us to measure changes in performance (particularly, standard Stroop interference) from the pre-double-response blocks to the post-double-response blocks.

Each standard and double-response block had an equal number of conflict and congruent trials randomly permuted within a 24-trial block.

4.1.3. Procedure and apparatus

The procedure and apparatus were the same as in Experiment 1 except that the experimental software would not continue with the next trial until a response of each type had been given in the correct order.

4.1.4. Results and discussion

This section is divided into three parts corresponding to each of three phases of the experiment: before, during, and after the double-response blocks in which the experimental manipulation occurred. Note that the key model prediction involves a change in Stroop interference across the experiment; this is tested in the subsection on performance *after* the double-response blocks.

4.1.4.1. Performance before the double-response blocks. Figure 5a displays the reaction time and error data for both experimental groups for the three block types occurring before the double-response blocks. Participants in both groups were fast and highly accurate in respond-

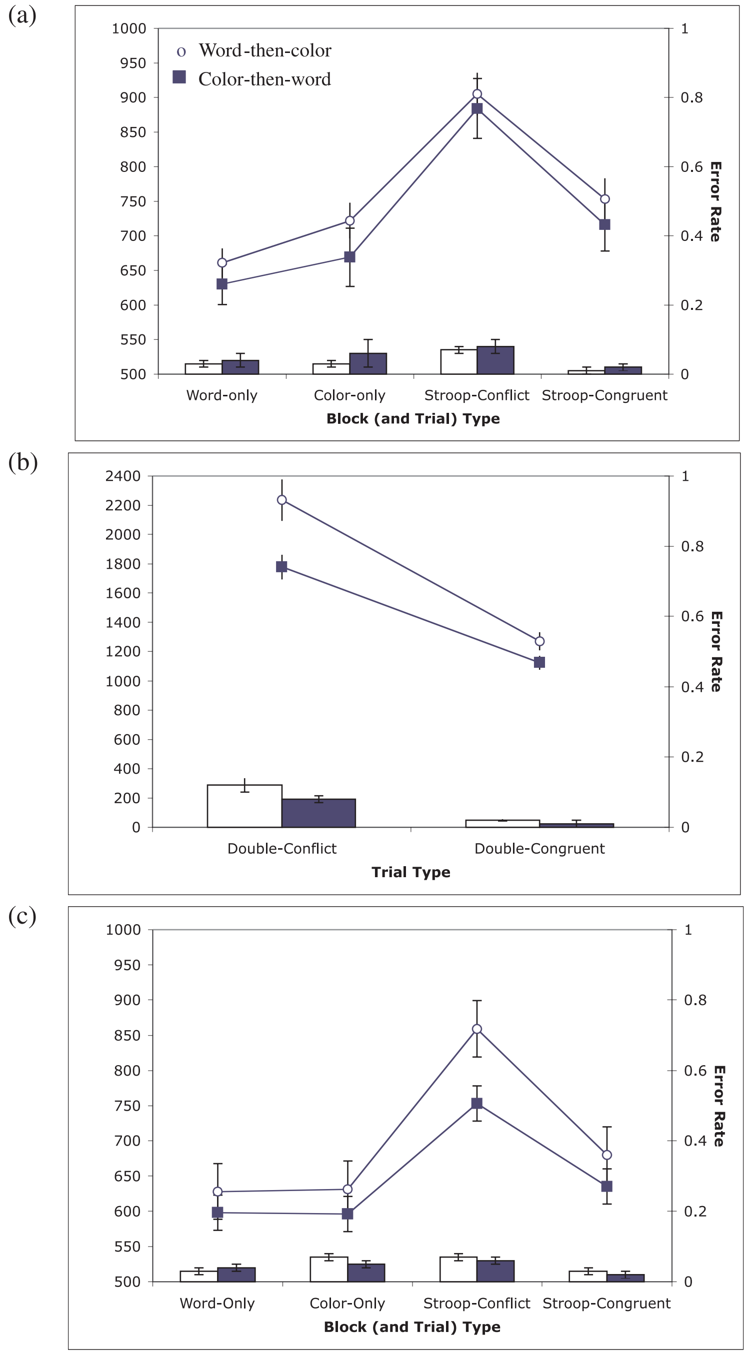

Fig. 5. Experiment 2 reaction times (lines) and error rates (bars) *before* the double-response blocks (panel a), *during* the double-response blocks (panel b), and *after* the double-response blocks (panel c). The left vertical axis corresponds to the reaction-time data, and the right vertical axis corresponds to the error rate data. Open circles and open bars refer to word-then-color condition; filled squares and filled bars refer to color-then-word condition.

Table 3
ANOVA results for Experiment 2, pre-double-response blocks

Task: Dependent Measure	Factor Analyzed	*F*	*MSE*	*p* value
Word-only: RT	Group	$F(1, 39) = .77$	12700	*ns*
Word-only: Errors	Group	$F(1, 39) = .35$	.003	*ns*
Color-only: RT	Group	$F(1, 39) = 1.15$	24200	*ns*
Color-only: Errors	Group	$F(1, 39) = .80$	.014	*ns*
Standard Stroop: RT	Group	$F(1, 39) = .37$	44400	*ns*
	Trial type	$\mathbf{F(1, 39) = 65}$	8050	**.0001**
	Group × Trial type	$F(1, 39) = .17$	8050	*ns*
Standard Stroop: Errors	Group	$F(1, 39) = 1.0$	.003	*ns*
	Trial type	$\mathbf{F(1, 39) = 29.5}$	.002	**.0001**
	Group × Trial type	$F(1, 39) = .02$	.002	*ns*

Note. Boldface indicates significant effect. ANOVA = analysis of variance; *MSE* = mean square error; RT = reaction time; *ns* = nonsignificant.

ing to word-only trials, relatively fast and highly accurate in responding to color-only trials, and showed significant Stroop interference in both reaction time (approximately 150 msec) and error rate (6%) for standard Stroop trials. This is consistent with other reports of Stroop interference.

For the pre-double-response blocks, performance between the two experimental conditions did not differ; that is, random assignment to the two conditions was satisfactory. See Table 3 for F values of each ANOVA, conducted as a one-way analysis (with experimental group as the single factor) for the word-only and color-only tasks and as a two-way mixed analysis (Experimental group × Trial type) for the standard Stroop task. Of all these analyses, the only significant effect was the main effect of trial type (i.e., significant Stroop interference) for both the reaction time and error data in the standard Stroop blocks.

4.1.4.2. Performance during the double-response blocks. Figure 5b shows total reaction times and error rates for both experimental groups for the double-response Stroop trials. Note that, for the double-response task, a trial is counted as having an error if *either* response was incorrect. These data were submitted to a two-way, mixed ANOVA, with response order as the between-subject factor and trial type as the within-subjects factor. Results indicated that the word-then-color group took longer, $F(1, 39) = 6.9$, $p < .05$, and made more errors, $F(1, 39) = 6.0$, $MSE = .002$, $p < .05$, than the color-then-word group. Both groups showed a significant effect of trial type—that is, significant Stroop interference—in reaction times, $F(1, 39) = 204$, $p < .001$, and errors, $F(1, 39) = 97.1$, $MSE = 4.0$, $p < .001$. Moreover, the Group × Trial-type interaction showed that the size of Stroop interference was larger for the word-then-color group than for the color-then-word group for reaction times, $F(1, 39) = 7.6$, $p < .01$, and marginally for errors, $F(1, 39) = 4.03$, $MSE = .001$, $p = .05$.

These between-group differences highlight the difficulty of the word-then-color response order and argue against different speed-accuracy trade-offs between groups. Finally, it should be noted that, for both groups, total reaction times for double-response conflict trials were more than twice those for standard Stroop conflict trials.

4.1.4.3. Performance after the double-response blocks. Figure 5c presents the reaction time and error data for both experimental groups for the three postmanipulation block types. Recall that this model's prediction is that practicing color-then-word response order in the double-response trials will increase the utility associated with color-processing production rules enough to favor processing the color dimension. This strategy shift toward color-naming first will, in turn, reduce Stroop interference in the subsequent standard Stroop blocks. Conversely, practicing word-then-color response order in the double-response trials will accentuate participants' preexisting high utility associated with word-processing production rules and hence maintain or slightly increase Stroop interference in the subsequent standard Stroop blocks.

To test these hypotheses regarding changes in Stroop interference, the data from the standard Stroop blocks occurring before and after the double-response blocks were submitted to a $2 \times 2 \times 2$ mixed ANOVA, with factors being group (color-then-word vs. word-then-color), trial type (conflict vs. congruent), and phase (pre- vs. postmanipulation). For the error data, there was only one significant effect, namely that more errors were committed on conflict trials, $F(1, 39) = 31.1$, $MSE = .003$, $p < .001$.

There were two main effects and one interaction in the reaction-time data. First, participants' reaction times were faster in the postmanipulation phase than in the premanipulation phase, $F(1, 39) = 21.4$, $MSE = 13000$, $p < .001$. This is consistent with a general speedup from increased familiarity with the experimental setup and with the key mapping for the colors. Second, participants were slower on conflict trials, $F(1, 39) = 95$, $MSE = 10280$, $p < .001$. This result points again to significant Stroop interference. The 3-way interaction of Group × Trial type × Phase reached marginal significance, $F(1, 39) = 3.2$, $MSE = 4750$, $p = .08$. This interaction is consistent with the hypotheses stated earlier about *different* changes in Stroop interference for the two groups. To test whether this interaction is of the form predicted by our hypotheses, we computed each participant's change in Stroop interference between the premanipulation and postmanipulation phases. That is, we created a new variable: (postStroopRT[conflict]—postStroopRT[congruent])—(preStroopRT[conflict]—preStroopRT[congruent]). Specifically, our prediction is to see a reduction in Stroop interference for the color-then-word group and a slight increase in (or maintenance of) Stroop interference for the word-then-color group. The data fit these predictions precisely: The color-then-word group on average showed a 50 (±27) msec reduction in Stroop interference, and the word-then-color group showed a 27 (±33) msec increase in Stroop interference. A two-sample *t* test comparing the conditions' change in Stroop interference showed a significant difference in the predicted direction, $t(39) = -1.8$, $p < .05$ (one-sided). In addition, comparing each group's average change in Stroop interference to zero revealed that the color-then-word group's change showed a significant reduction $t(20) = -1.8$, $p < 05$ (one-sided), and the word-then-color group's change was not significantly different from zero $t(19) = 0.814$.

Although the double-response response-order manipulation was predicted to impact Stroop interference, it was not predicted to impact the color-only or word-only blocks where there is no competing second dimension. Thus, the color-only and word-only blocks serve as control tasks. ANOVAs were performed on both block types' reaction time and error data, with group (word-then-color vs. color-then-word) and experimental phase (pre-double-blocks vs. post-double-blocks) as factors. In all cases the interaction was not significant, as expected. For reaction times, both block types showed significant speedup from pre- to postmanipulation:

$F(1, 39) = 30.7$, $MSE = 4500$, $p < .001$, for color-only blocks and $F(1, 39) = 12.3$, $MSE = 1700$, $p < .005$, for word-only blocks. This is likely the same generic speedup that we found for the Stroop blocks. It is somewhat surprising that there would be speedup in the word-only blocks because reading is so highly practiced, but the speedup was only on the order of 30 msec. For the error data in the color-only and word-only blocks, there were no significant effects.

4.1.5. Modeling change in Stroop interference

The same model used to produce the model fits discussed in the introduction was used to fit these data. The only difference is that for this experiment the model needed one extra parameter to initialize utility learning. Thus, with three free parameters, the model was fit to 18 data points, minimizing the mean deviation between model and data. Twelve of these data points were raw Stroop reaction times (see Figure 6), and the other six were conflict—congruent reaction-time differences, included to emphasize interference effects in optimizing the model fit. This model fit had a mean deviation of 62.1 msec and $R^2 = .97$.

Figure 6 shows that the model is able to capture several effects in the data: Conflict trials are slower than congruent trials, standard Stroop trials are faster than double-response trials, double-response word-first responses are slower than double-response color-first responses (although the model underestimates this effect), and response order in the double-response blocks impacts standard Stroop interference. This last result is the key finding of Experiment 2, namely that the size of Stroop interference decreased for the color-then-word group and (slightly) increased for the word-then-color group. The model similarly showed an initial Stroop interference effect of 176 msec, which decreased to 105 msec in the color-then-word condition and increased to 214 msec in the word-then-color condition.

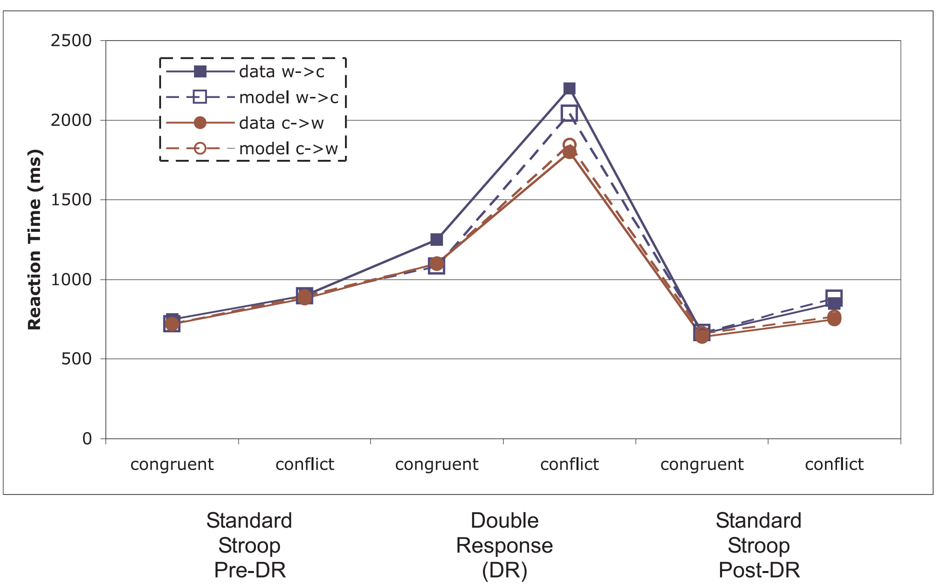

Fig. 6. Model fit to key trial types in Experiment 2.

These model results stem from the model's ability to initiate processing with either the word or color dimension and from its ability to shift its bias from one response order to another, based on their learned utilities. For example, because of its high initial utility for word reading, the model tends to process the "correct" stimulus dimension first in the word-then-color condition but the "wrong" one in the color-then-word condition. This experience updates the utility of production rules for processing each stimulus dimension, with more utility change occurring in the color-then-word condition. This bias carries forward to the post-double-response standard Stroop blocks, producing reduced Stroop interference in the color-then-word case and slightly increased Stroop interference in the word-then-color case.

5. Experiment 3

Experiment 2 shows that instructions specifying a particular response order in the double-response Stroop task can experimentally impact the size of participants' Stroop interference effect. Experiment 3 aimed to replicate this finding and expand it to a slightly different situation. The only procedural change from Experiment 2 to 3 was that all participants were asked to produce *both* responses manually. That is, the mapping of stimulus dimension (word vs. color) to response was not a mapping to modality (verbal vs. manual) but rather a mapping to a particular row of keys. This had the advantage of making the two responses more comparable in response times (i.e., simple horse-race accounts would be insufficient). Further, this offered a different approach to the response-mapping compatibility issues that surfaced in Experiment 1: Instead of eliminating these issues (as in Experiment 2 by studying only the high-compatibility response mapping), Experiment 3 made both responses manual. In effect, this should make the task somewhat more difficult than the compatible response mapping (Experiments 1 and 2) but less difficult than the incompatible response mapping (Experiment 1). Recall that this model predicts that an increase in task difficulty, especially when it increases response times, would lead to a larger change in Stroop interference after the double-response manipulation. Thus, the qualitative prediction going into this experiment is the same as Experiment 2: The color-then-word condition should show a reduction in Stroop interference after double-response blocks. Moreover, because of Experiment 3's more difficult version of the double-response Stroop task, this should be somewhat magnified relative to Experiment 2.

The only study-design change in Experiment 3 is that the second of the two experimental conditions was no longer another version of the double-response task but rather an equal number of trials of standard Stroop. This condition offers a new control for the color-then-word double-response condition. A reasonable expectation would be that more practice at standard Stroop in the manipulation phase would be the best way to reduce Stroop interference—even better than double-response Stroop. This model, in contrast, predicts reduction in Stroop interference for both groups, but more for the color-then-word double-response condition.

5.1. Methods

5.1.1. Participants

The participants in this experiment were 53 Carnegie Mellon University undergraduates who received course-related credit for participating.

5.1.1.1. Design. The design of this experiment included one between-subject manipulation, specifying which task participants performed in the middle blocks of the experiment: either double-response Stroop (in the color-then-word order) or standard Stroop.

5.1.2. Procedure, apparatus, and analysis

Except for the differences noted previously, the methods were the same as in Experiment 2. Data analyses focus on reaction times and use the same analysis procedure as Experiments 1 and 2.

5.1.3. Results and discussion

This section is divided into three parts corresponding to the three phases: before, during, and after the experimental manipulation. Note that, as with Experiment 2, the critical model prediction involves a change in Stroop interference from pre- to postmanipulation blocks; this is tested in the section on performance *after* the manipulation.

5.1.3.1. Performance before the experimental manipulation. Figure 7a displays reaction times for both groups for the three premanipulation block types: word-only, color-only, and standard Stroop. In the case of the standard Stroop blocks, reaction times are plotted separately for conflict and congruent trials. Participants in both groups took approximately 650 msec for the word-only and color-only tasks, took longer for the standard Stroop task, and showed a significant Stroop interference effect (approximately 150 msec), $F(1, 51) = 100.1$, $MSE = 6528$. This last finding is consistent with Stroop interference effects found in other reports, including Experiment 2. The word-only reaction time may seem longer than typical results in the literature, but recall that in this experiment participants were learning new key mappings for manual responding. Most other reports of word-only reaction times involve highly practiced verbal responses. Also note that, as in Experiment 2, the expectation was that performance in these premanipulation blocks should not differ between conditions. This expectation was supported for all three tasks, word-only: $F(1, 51) = 1.80$, $MSE = 17100$; color-only: $F(1, 51) = 0.99$, $MSE = 13690$; standard Stroop: $F(1, 51) = 2.2$, $MSE = 42400$.

5.1.3.2. Performance during the experimental manipulation. During the experimental manipulation blocks, participants were performing either the double-response Stroop (responding to the color *first*) or the standard Stroop task (i.e., responding to the color *only*). Therefore, it is not surprising that the double-response group had much longer reaction times, $F(1, 51) = 47.5$, $MSE = 341280$. It is also not surprising that there was a significant trial-type effect, with conflict trials taking longer than congruent trials, $F(1, 51) = 43.2$, $MSE = 87300$. Similar to the result in Experiment 2, the interaction of these two effects was also significant, with the larger trial-type effect exhibited by the double-response group, $F(1, 51) = 27.9$, $MSE = 87300$. Again,

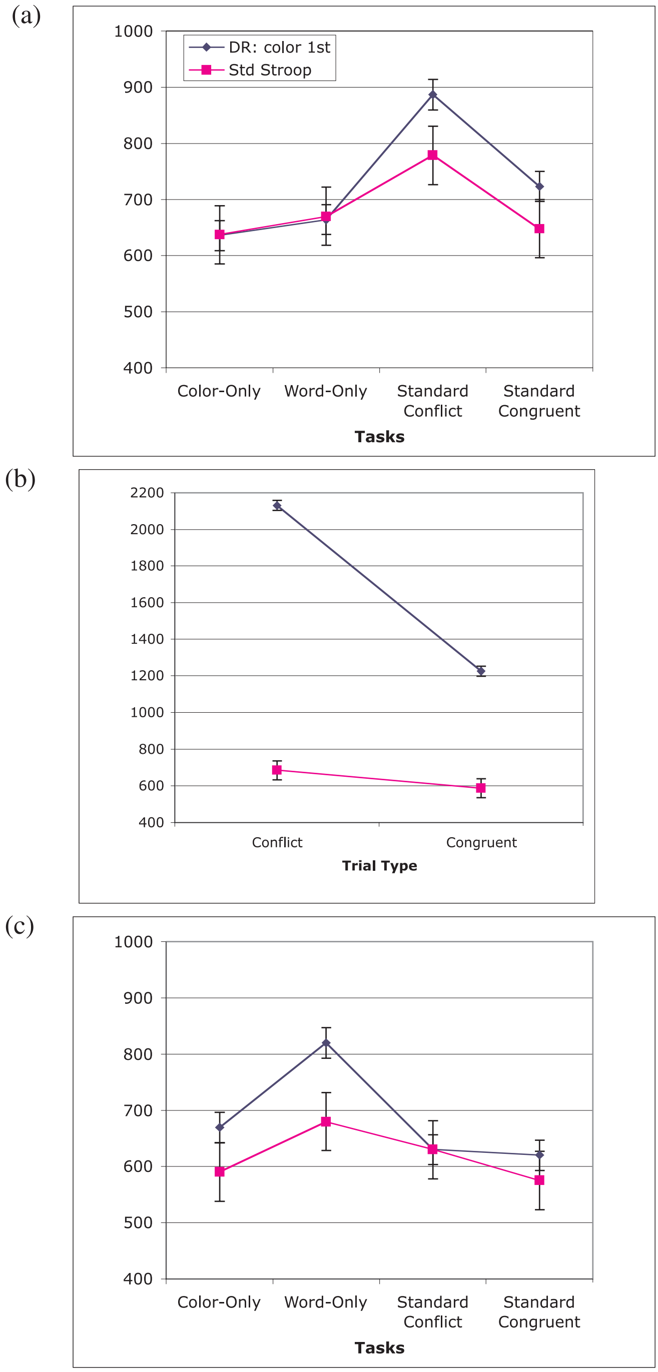

Fig. 7. Experiment 3 reaction-time data (a) *before* the experimental manipulation, (b) *during* the experimental manipulation, and (c) *after* the experimental manipulation.

this interaction is likely an example of Stroop interference scaling with the absolute magnitude of reaction times. In fact, looking at the data in Figure 7b, it is important to notice the scale of the *y* axis to realize that the difference between conflict and congruent trials for the standard Stroop participants is actually a full 100 msec; this effect appears small mainly because of the much larger double-response reaction times.

5.1.3.3. Performance after the experimental manipulation. Figure 7c presents the reaction times for both groups for the three postmanipulation block types: word-only, color-only, and standard Stroop. As in Experiment 2, the key model prediction involves testing whether the experimental manipulation had an impact on Stroop interference. In Experiment 2, that manipulation involved different response orders for the double-response Stroop, and the finding was that the color-first response order significantly reduced Stroop interference, whereas the word-first response order slightly amplified Stroop interference. In Experiment 3, the color-first double-response Stroop task was again expected to reduce Stroop interference. Here, however, the comparison condition was the standard Stroop task, another task where processing color first has increased utility. Thus, the model predictions for this experiment were that (a) both tasks will lead to a reduction in Stroop interference, and (b) this reduction will be greater for the double-response condition. These predictions were tested by a 2 × 2 mixed ANOVA on the size of Stroop interference, where the between-subject factor was the experimental manipulation (double-response or standard Stroop) and the within-subjects factor was phase of the experiment (pre- or postmanipulation). Supporting prediction (a), this analysis showed a significant main effect of phase of the experiment (pre- vs. postmanipulation), such that Stroop interference effects were smaller after either task manipulation, $F(1, 51) = 26.9$, $MSE = 7276$. Supporting prediction (b), these data showed an interaction between condition and phase of the experiment. As predicted, greater reduction in Stroop interference occurred for the double-response group, $t(51) = 1.77$, $p < .05$. (Note that this analysis is equivalent to the *t* test computed in Experiment 2, comparing pre- to postmanipulation changes in Stroop interference between conditions.)

Analyzing participants' reaction times in the word-only and color-only blocks before versus after the manipulation, only two significant effects were observed. The first of these was a significant increase in word-only reaction times, $F(1, 50)$, 14.2, $MSE = 9684$. This main effect, however, should be interpreted in light of its interaction with condition, $F(1, 50) = 5.6$, $MSE = 9684$. Specifically, this interaction showed that the double-response group showed a greater increase in word-only reaction times. These results are consistent with the utility of word reading decreasing relative to color naming for both conditions but more so in the color-then-word double-response case.

5.1.4. Modeling change in Stroop interference

The model that was fit to Experiment 2 was slightly adjusted to accommodate the main procedural change in Experiment 3, namely, that all participants were responding manually to both stimulus dimensions. This change was incorporated into the model by no longer considering all the word-reading association chunks to have high activation. In fact, the word-association chunks were set to have the same activation as the color-association chunks, and this activation value was a free parameter in the model fit. Although this activation value was adjusted to account for the new (manual) response mapping, the production rules for word processing and color processing started the experiment as they had in Experiment 2, with a utility-based bias toward processing the word first. From this starting point, the utility values of the two production rules shifted, based on the experience participants received in the two conditions. For the double-response condition, the utility values shifted as they had for the corresponding (color-then-word) condition in Experiment 2. For the standard Stroop condition, a single parameter was applied to account for the utility learning among participants in that group. All other parameters of the model were kept the same as the fit from Experiment 2.

Figure 8 shows the model fit for both conditions, for conflict and congruent trials across the three phases of the experiment—premanipulation standard Stroop trials, manipulation trials (either double-response or standard Stroop), and postmanipulation standard Stroop trials. This fit has $R^2 = .99$ and a mean deviation of 43.8 msec. Notice the data for the two groups in the premanipulation phase are numerically different. Although this is not a statistically reliable difference, the sampling effect of slightly faster reaction times among those in the standard Stroop condition is not accounted for by the model.

The model captures the four key results here. First, it captures the approximately 170 msec premanipulation Stroop effect as it had in Experiment 2. Second, it captures the very large Stroop effect in the double-response task as it had in Experiment 2. Third, it captures a gradual shrinking of the Stroop effect for the standard Stroop condition across the experiment by its shifting of the utility values for reading and color naming. Fourth, it captures a more substantial reduction in the postmanipulation Stroop effect for the double-response condition because of a more substantial shift in utility values. As discussed earlier, the utility value for color naming relative to word reading shifts more in the double-response condition because there is a greater benefit to processing the color first in this condition.

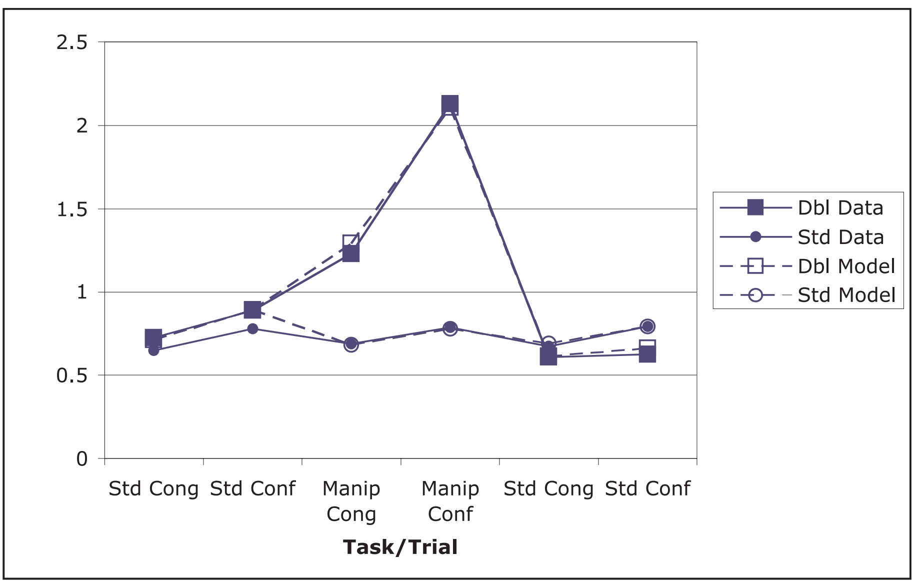

Fig. 8. Model fit to key trial types in Experiment 3.

6. General discussion

This article presented new modeling and empirical work regarding Stroop effects. The main idea behind this model is that people can approach Stroop tasks by processing the color *or* the word first and that this implicit choice reflects a learned (although likely still implicit), strategy choice based on the relative utility of word reading versus color naming. A model implementing this idea within the ACT–R architecture accounts for basic Stroop effects and several new Stroop results.

Specifically, this model made several novel predictions regarding particular Stroop-related experiences that could change the size of participants' Stroop interference. First, the model predicted a significant *decrease* in standard Stroop interference after double-response Stroop trials under color-then-word response-ordering instructions. Second, the model predicted a moderate *increase* in standard Stroop interference after double-response Stroop trials under word-then-color instructions. These predictions were confirmed in an observational study (Experiment 1) and under an experimental manipulation (Experiment 2). Third, the model predicted that additional practice at the standard Stroop task would also reduce the size of Stroop interference but not as much as experience at responding color-then-word in the double-response task. This prediction was also supported (Experiment 3).

6.1. Model implications

Utility learning is the key mechanism that this model uses to explain the observed Stroop effects. In this model, utility is a measure of how effective a production rule has been, averaged over the instances in which it was applied. Specifically, utility is computed in terms of successes, failures, and costs. This emphasis on the outcome of a production rule's firing over its sheer amount of practice harkens to the Law of Effect versus the Law of Practice (Thorndike, 1927, 1932) The Law of Effect states that actions that are effective (e.g., in achieving goals) will be repeated in the future. This is what happens with the color-naming production rule in some of the conditions investigated here, and when it is particularly effective, it becomes more likely to fire at the beginning of a trial, thus reducing subsequent Stroop interference. The insight of this model is that utility, a construct akin to effectiveness, is learned at the level of the production rule and that *relative* utility is what determines repeated future use. This is consistent with the results presented in this article and with past work (Lovett & Anderson, 1996; Singley & Anderson, 1989).

In contrast, the Law of Practice states that mere practice of an action leads to its repeated use. This view is more consistent with practice-based models of Stroop (e.g., Cohen et al., 1990). It predicts that practicing a skill—regardless of the effectiveness of that skill—will lead to more use of the skill. This is not what was found in these studies. For example, the two response orders for the double-response Stroop task had the same amount of practice with color naming and word reading but produced different changes—a decrease versus an increase—in Stroop interference depending on the response order. Practice-based models would predict a reduction in interference under both conditions.

This model includes both learning based on effectiveness and learning through practice. As an illustration of how it differs from a practice-based model, we fitted a practice-only variant of this model to the data from Experiment 2. Specifically, we disabled this model's utility-learning features but retaining its practice-based features, that is, the strengthening of color-association and word-association chunks. This is a reasonable practice-based model and is very similar to earlier ACT–R models of Stroop. Although this model produces an overall decrease in reaction times with practice that is consistent with the data, even the best-fitting version of this model fails to produce different-sized reductions in Stroop interference for the two response-order conditions: Its initial Stroop interference of 180 msec decreases to 164 msec for both conditions. The fit of this practice-only model (with mean deviation = 116.6 msec, $R^2 = .94$) is significantly worse than the fit of this model presented earlier ($p < .05$). This failure to adequately account for changes in Stroop interference is also consistent with results from practice-based models in the literature. For example, Figure 12 in Cohen et al. (1990) shows model performance that decreases its reaction times with practice, but that model produces changes in Stroop interference that differ systematically from what is observed in the MacLeod and Dunbar (1988) study. In contrast, this model was able to fit the changing nature of these Stroop interference data quite well (see Figure 2).

Botvinick et al. (2001) extended the practice-based model of Cohen et al. (1990) by adding another layer that can adjust its task-related control. This model posits that greater control (from task control nodes) is exerted when there is more "crosstalk" in the system, where crosstalk is a model-based measure of conflict in the network. With this extra component, the model can account for data from Tzelgov et al. (1992) that show a reduction in the size of Stroop interference as a function of proportion of conflict trials. As mentioned in Section 2.3, this model can also address this effect because the key production rules' utility values change as a result of experience (e.g., lower accuracies and slower responses with a greater proportion of conflict trials). In the Botvinick et al. model, control is modulated by crosstalk. As a rule of thumb, crosstalk tends to be high when performance is low, and low when performance is high. This is consequently a proxy for processing effectiveness, which we see as introducing an element of utility into a practice-based account.

Similarly, one could conceive of a practice-based model variant in which a greater learning benefit is gained during double-response Stroop from practice on the first-processed dimension than the second-processed dimension (e.g., due to some attentional bias). This account would be consistent with the results in Experiment 2 in that it would produce greater benefit for color processing in the color-then-word condition and hence would also produce subsequent reduced Stroop interference. But this account also predicts a greater reduction in Stroop interference after standard Stroop practice—where color processing would get full attention—compared to double-response practice. This is inconsistent with the results from Experiment 3.

6.2. *Learning new production rules*

One feature of this model that plays a role in the predictions discussed throughout this article is that it uses many of the same production rules in the double-response and standard Stroop tasks. That enables the utility-learning effect gained from the double-response trials to carry over to strategy choices in the standard Stroop task. Employing the same production rules in both tasks is parsimonious in that the double-response task is simply a combination of standard

Stroop tasks. Moreover, the model's fit to the data in Experiments 2 and 3 supports this knowledge representation. An interesting question, however, is whether people could learn a new production rule for the double-response task given sufficient practice. If so, any utility learning in the double-response task would apply to this new production and would not carry over to the standard Stroop task. This idea suggests that giving participants more and more experience at the double-response Stroop task could eventually lead to a *reduced* impact on their standard Stroop interference effect. This idea has been explored in Anderson et al. (2004) and is at the heart of empirical questions currently being explored by Hazeltine and colleagues (Hazeltine, Teague, & Ivry, 2002).

What can we say about reducing Stroop interference?

Because the Stroop effect is so robust, any result that shows it being reduced or eliminated warrants attention. Several manipulations have been shown to reduce Stroop interference, and some of these have been highlighted in this article. First, most simply, additional practice at the standard Stroop task does reduce Stroop interference, albeit gradually. Experiment 3 of this article showed a reduction in Stroop interference from approximately 160 msec down to 100 msec. This occurred with standard Stroop trials distributed equally between conflict and congruent types but where word *and* color responses were made manually. Note that the verbal-response Stroop would produce a smaller reduction in interference because verbal responses are faster than manual responses and thus leave less room for utility changes. Other results that show changes in Stroop interference were discussed earlier. These include the uneven distribution of trial types (Stroop, 1935; Tzelgov et al., 1992) and the learning of a new competing task (MacLeod & Dunbar, 1988).

It should be emphasized that these approaches to reducing Stroop interference require more than simply giving more standard Stroop practice. In each case, there was something special about the Stroop task participants performed: unusual response mode, biased trial-type composition, or novelty of the tasks involved. A reasonable heuristic seems to be that the less "standard" the Stroop task being practiced, the greater impact it will have on reducing Stroop interference. From the model's point of view, this is explained by the fact that standard Stroop task practice does not allow for a very large utility shift between word reading and color naming. Task variants that place a higher cost on word reading (or conversely, a lower cost on color naming) will show the largest change in utility and hence the greatest possible reduction in Stroop interference.

References

Altmann, E. M., & Davidson, D. J. (2001). An integrative approach to Stroop: Combining a language model and a unified cognitive theory. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd annual conference of the Cognitive Science Society* (pp. 21–26). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111,* 1036–1060.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108,* 624–652.

Cheesman, J., & Merikle, P. M. (1984). Priming with and without awareness. *Perception and Psychophysics, 36,* 387–395.

Chrisman, S. D. (2001). Individual differences in Stroop and local-global processing: A possible role of interhemispheric interaction. *Brain & Cognition, 45,* 97–118.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97,* 332–361.

Cohen, J. D., & Huston, T. A. (1994). Progress in the use of interactive models for understanding attention and performance. In C. Umilta, & M. Moscovitch (Eds.), *Attention and performance 15: Conscious and nonconscious information processing* (pp. 453–476). Cambridge, MA: MIT Press.

Comalli, P. E., Jr., Wapner, S., & Werner, H. (1962). Interference effects of Stroop color-word test in childhood, adulthood, and aging. *Journal of Genetic Psychology, 100,* 47–53.

Dalrymple-Alford, E. C., & Azkoul, J. (1972). The locus of interference in the Stroop and related tasks. *Perception & Psychophysics, 11,* 385–388.

Dalrymple-Alford, E. C., & Budayr, B. (1966). Examination of some aspects of the Stroop color-word test. *Perceptual & Motor Skills, 23*(3, Pt. 2), 1211–1214.

Glaser, M. O., & Glaser, W. R. (1982). Time course analysis of the Stroop phenomenon. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 875–894.

Greenwald, A. G. (1972). On doing two things at once: Time sharing as a function of high ideomotor compatibility. *Journal of Experimental Psychology, 94,* 52–57.

Hazeltine, E., Teague, D., & Ivry, R. (2002). Simultaneous dual-task performance reveals parallel response selection after practice. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 527–545.

Kane, M., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General 132,* 47–70.

Kerns, J. G., Cohen, J. D., MacDonald, A. W. III, Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science, 303,* 1023–1026.

Klein, G. S. (1964). Semantic power measured through the interference of words with color-naming. *American Journal of Psychology, 77,* 576–588.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(1), 1–38.

Lindsay, D. S., & Jacoby, L. L. (1994). Stroop process dissociations: The relationship between facilitation and interference. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 219–234.

Logan, G. D. (1980). Attention and automaticity in Stroop and priming tasks: Theory and data. *Cognitive Psychology, 12,* 523–553.

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Memory and Cognition, 7,* 166–174.

Logan, G. D., & Zbrodoff, N. J. (1982). Constraints on strategy construction in a speeded discrimination task. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 502–520.

Logan, G. D., Zbrodoff, N. J., & Williamson, J. (1984). Strategies in the color-word Stroop task. *Bulletin of the Psychonomic Society, 22,* 135–158.

Lovett, M. C. (1998). Choice. In J. R. Anderson, & C. Lebiere (Eds.), *The atomic components of thought* (pp. 255–296). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Lovett, M. C. (2002). Modeling selective attention: Not just another model of Stroop. *Cognitive Systems Research, 3,* 67–76.

Lovett, M. C., & Anderson, J. R. (1996). History of success and current context in problem solving. *Cognitive Psychology, 31,* 168–217.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109,* 163–203.

MacLeod, C. M. (1998). Training on integrated versus separated Stroop tasks: The progression of interference and facilitation. *Memory & Cognition, 26,* 201–211.

MacLeod, C. M., & Dunbar, K. (1988). Training and Stroop-like interference: Evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 126–135.

MacLeod, C. M., & MacDonald, P. A. (2000). Inter-dimensional interference in the Stroop effect: Uncovering the cognitive and neural anatomy of attention. *Trends in Cognitive Sciences, 4,* 383–391.

Neill, W. T. (1978). Decision processes in selective attention: Response priming in the Stroop color-word task. *Perception and Psychophysics, 23,* 80–84.

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

Phaf, R. H., van der Heijden, A. H., & Hudson, P. T. (1990). SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology, 22,* 273–341.

Roelofs, A. (2002). Control of language: A computational account of the Stroop asymmetry. In N. Taatgen & J. Aasman (Eds.), *Proceedings of the Third International Conference on Cognitive Modeling* (pp. 234–241). Veenendaal, The Netherlands: Universal Press.

Roelofs, A. (2003). Goal-referenced selection of verbal action: Modeling attentional control in the Stroop task. *Psychological Review, 110,* 88–125.

Roelofs, A., & Hagoort, P. (2002). Control of language use: Cognitive modeling of the hemodynamics of Stroop task performance. *Cognitive Brain Research, 15*(1), 85–97.

Schiller, P. H. (1966). Developmental study of color-word interference. *Journal of Experimental Psychology, 72,* 105–108.

Schooler, C., Neumann, E., Caplan, L. J., & Roberts, B. R. (1997). A time course analysis of Stroop interference and facilitation comparing normal individuals and individuals with schizophrenia. *Journal of Experimental Psychology: General, 126,* 19–36.

Schweickert, R. (1978). A critical path generalization of the additive factor method: Analysis of a Stroop task. *Journal of Mathematical Psychology, 18,* 105–139.

Shimada, H., & Nakajima, Y. (1991). Double response to Stroop stimuli. *Perceptual & Motor Skills, 72,* 571–574.

Siegler, R. S. (1991). Strategy choice and strategy discovery. *Learning and Instruction, 1*(1), 89–102.

Siegler, R. S. (1996). *Emerging minds.* New York: Oxford University Press.

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill.* Cambridge, MA: Harvard University Press.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643–662.

Sugg, M. J., & McDonald, J. E. (1994). Time course of inhibition in color-response and word-response versions of the Stroop task. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 647–675.

Thorndike, E. L. (1927). The law of effect. *American Journal of Psychology, 39,* 212–222.

Thorndike, E. L. (1932). *The fundamentals of learning.* New York: Columbia University.

Tzelgov, J., Henik, A., & Berger, J. (1992). Controlling Stroop effects by manipulating expectations for color words. *Memory & Cognition, 20,* 727–735.

Virzi, R. A., & Egeth, H. E. (1985). Toward a translational model of Stroop interference. *Memory & Cognition, 13,* 304–319.