Psychological Review
2016, Vol. 123, No. 3, 255–290

0033-295X/16/$12.00 http://dx.doi.org/10.1037/a0039979

An Integrated Utility-Based Model of Conflict Evaluation and Resolution in the Stroop Task

Adam Chuderski
Jagiellonian University

Tomasz Smolen
Pedagogical University of Krakow

Cognitive control allows humans to direct and coordinate their thoughts and actions in a flexible way, in order to reach internal goals regardless of interference and distraction. The hallmark test used to examine cognitive control is the Stroop task, which elicits both the weakly learned but goal-relevant and the strongly learned but goal-irrelevant response tendencies, and requires people to follow the former while ignoring the latter. After reviewing the existing computational models of cognitive control in the Stroop task, its novel, integrated utility-based model is proposed. The model uses 3 crucial control mechanisms: response utility reinforcement learning, utility-based conflict evaluation using the Festinger formula for assessing the conflict level, and top-down adaptation of response utility in service of conflict resolution. Their complex, dynamic interaction led to replication of 18 experimental effects, being the largest data set explained to date by 1 Stroop model. The simulations cover the basic congruency effects (including the response latency distributions), performance dynamics and adaptation (including EEG indices of conflict), as well as the effects resulting from manipulations applied to stimulation and responding, which are yielded by the extant Stroop literature.

Keywords: Stroop task, cognitive control, conflict monitoring, response utility, reinforcement learning

Cognitive control (also called *executive control*) allows humans to direct and coordinate their own thoughts and actions in a flexible and novel way in order to reach adopted goals, even in the face of conflicting stimulation and strongly learned but inadequate response tendencies. For example, when choosing between preparing for a boring but important exam and going for a pleasant but distracting party, most people will require mental effort and reflection to reject the party, and some people fail to do that. Moreover, due to cognitive control, high-level behavioral plans can be transformed into adequate, albeit never learned, sequences of thoughts and actions, necessary for our effective coping with novel and demanding situations. The important role of cognitive control for coherent behavior can easily be seen in cases when such control has been disrupted (e.g., due to mental overload, tiredness, aging, neurological diseases, etc.), resulting in no longer being able to inhibit prepotent but improper responses, prevent perseveration, overcome distraction and interference, switch between tasks, and organize actions into meaningful sequences (Kimberg & Farah, 1993; Norman, 1981; Reason, 1990; Shallice & Burgess, 1993).

One of the fundamental aims of psychology, neuroscience, and cognitive science is to understand how the mind/brain can internally control its own cognitive processing, without positing any vague, homuncular constructs like *will*, *person*, or *self* (Monsell & Driver, 2000). Recent contributions to the problem of cognitive control have begun to successfully address questions of *what*, in fact, controls cognition and precisely *how* this control is exerted. First, cognitive control does not seem to be a function of one dedicated cognitive subsystem, but likely emerges from complex interactions between diverse neurocognitive mechanisms, which can be algorithmically specified in terms of computational models (Braver, 2012; Gruber & Goschke, 2004; Hazy, Frank, & O'Reilly, 2007; Verguts & Notebaert, 2008). Importantly, a special role in integrating the interplay of all these mechanisms has been attributed to *goal* representations (see Austin & Vancouver, 1996), which help to select (Anderson & Lebiere, 1998; Gollwitzer, 1993; Miller, Galanter, & Pribram, 1960; Newell & Simon, 1972), or at least guide (Cohen, Dunbar, & McClelland, 1990; Norman & Shallice, 1986) goal-relevant processes, while suppressing other conflicting processes (Hasher, Lustig, & Zacks, 2007; Munakata et al., 2011; Nigg, 2000).

Second, progress has been made in the understanding of the mechanisms yielding goal-directed behavior. Specifically, when, why, and how strongly goals regulate cognitive processing have been investigated, for instance, by proposing that cognitive control responds to the evaluated level of internal conflict in a cognitive system (Botvinick, Braver, Barch, Carter, & Cohen, 2001), the likelihood of negative events like incongruency, errors, or risks (Brown & Braver, 2005, 2008; Holroyd & Coles, 2002; Jiang, Heller, & Egner, 2014), and the discrepancy between the experienced and either intended (Scheffers & Coles, 2000) or predicted response outcomes (Alexander & Brown, 2011). Control seems to modulate processing only when it is really needed, but not when well-learned sequences of actions fully suffice for the successful

This article was published Online First January 11, 2016.

Adam Chuderski, Institute of Philosophy, Jagiellonian University; Tomasz Smolen, Department of Psychology, Pedagogical University of Krakow.

This work was sponsored by the National Science Centre of Poland (grant no. 2011/01/D/HS6/00467).

Correspondence concerning this article should be addressed to Adam Chuderski, Institute of Philosophy, Jagiellonian University, Grodzka 52, 31-044 Krakow, Poland. E-mail: adam.chuderski@gmail.com

performance. This seems to reflect the optimal control strategy, as prolonged and/or overacted cognitive control may be cognitively costly and inefficient (Botvinick et al., 2001; Taatgen, 2007).

Due to the complexity as well as covert nature of control processes, they have primarily been investigated using relatively simple, well-controlled experimental situations, instead of observing self-control in natural settings. The hallmark instance of such experimental tests was variants of the *Stroop task* (Stroop, 1935; see also MacLeod, 1991). In general, this task consists of presenting bivalent stimuli, which elicit the weakly learned but goal-relevant (i.e., nondominant), but also the strongly learned but goal-irrelevant (i.e., dominant) response tendencies, and requires people to follow the former while ignoring the latter. The present study is devoted to the understanding and explanation of the mechanisms of cognitive control exerted in the Stroop task, using a novel computational model. We start with a description of this task as well as the experimental effects that it yields. Then, after reviewing the existing Stroop models, we present our own model that integrates several prospective control mechanisms in a novel way. Finally, the model is corroborated in a series of 18 computer simulations.

The Stroop Task

The standard Stroop (1935) task requires naming the color of a word, which itself refers to a color. Two basic effects were observed in this task. The *interference effect* consists of increased response time (reaction time [RT]) in incongruent trials, when the color denoted by a word mismatches the ink color, compared with RT in neutral trials, when the color of an unrelated color string, like "XXXXX," has to be named. The *facilitation effect* consists of decreased RT in congruent trials, when a color and a word match, in comparison with neutral trials. The interference effect is consistently larger than is the facilitation effect (e.g., Dyer, 1971; see Simulation 1). Both of these *congruency effects*, as well as the related RTs, display *right-skewed distributions* (Mewhort, Braun, & Heathcote, 1992; Simulations 2 and 3). Early studies (e.g., Stroop, 1935; Thurstone, 1944) used lists of either neutral or incongruent stimuli (for a review see Jensen & Rohwer, 1966). Since Dalrymple-Alford and Budayr's (1966) study, the individual stimulus variant (i.e., one stimulus per trial) has started to dominate the field, allowing for fine-grained stimulus manipulations (also greatly supported by computerized methods; see MacLeod, 1991). Both vocal (i.e., saying the name of a color; Stroop, 1935) and manual (i.e., pressing keys associated with a color; White, 1969) response modes were used. The congruency effects were also observed in a variety of other congruency tests (named Stroop-like tasks), for instance, in naming pictures that included words (the *picture-word task*; Hentschel, 1973; Simulation 4), responding to symbols surrounded by other symbols (the flanker task; Eriksen & Eriksen, 1974), identifying the positions of words denoting positions (Seymour, 1973), counting numbers (Shor, 1971), and categorizing large objects comprised of smaller ones (Navon, 1977).

Although more recent studies indicated that the mechanisms responsible for interference in the standard color-word task may differ, to some extent, from interference generators in the Stroop-like tasks, like the flanker (Kornblum, Hasbroucq, & Osman, 1990) or the picture-word task (van Maanen, van Rijn, & Borst, 2009), a number of experimental effects have reliably been shown in various Stroop variants (see MacLeod, 1991). First, substantial *practice reverses interference* (MacLeod & Dunbar, 1988; Simulation 5). The interference effect increases with an increasing proportion of congruent trials in the stimuli sequence (the *proportion-congruent* and *item-specific proportion-congruent effects*; e.g., Jacoby, Lindsay, & Hessels, 2003; Logan & Zbrodoff, 1979; Tzelgov, Henik, & Berger, 1992; Simulations 6 and 7). The *effect of preceding trial's congruency* (the *Gratton effect*; Gratton, Coles, & Donchin, 1992; Simulation 8) is a decrease in the congruency effect for stimuli following incongruent stimuli, compared with stimuli following congruent stimuli. In EEG research, the interference effect is accompanied by a characteristic brain activity in the fronto-central cortex (the *N2/N450 and ERN waves*; Kopp, Rist, & Mattler, 1996; Simulations 9 and 10). A reduced but significant interference is noted when the presentation of each stimulus is temporally (the *stimulus onset asynchrony effect*, *SOA*; Glaser & Glaser, 1982; Simulation 11) or *spatially disintegrated* (Dyer, 1973; Kahneman & Henik, 1981; Simulation 12). Reduction in interference is also observed when incongruent stimuli belong to the same kind (e.g., *the smaller word-word and color-color interference*; Dallas & Merikle, 1976; Simulation 13), or when objects are not directly incongruent, but are, instead, just associated by *semantic gradient* (Klein, 1964; Simulation 14). The interference is larger when a word denotes a color associated with another potential response, compared with a color not included in the set of possible responses (the *response-set effect*; Proctor, 1978; Simulation 15). There is a small but often-significant *reverse interference effect* (e.g., Blais & Besner, 2006; Dunbar & MacLeod, 1984; Simulation 16), for instance when people are reading colored words. We refer in detail to these 16 crucial effects (as well as two novel findings) throughout the Simulations section. Moreover, many less important Stroop task findings were observed (MacLeod, 1991; Roelofs, 2003), which we address in the Comparison with Alternative Models of Stroop section.

Soon after the Stroop task was introduced, several theories attempted to explain the nature of the congruency effects it yields. One theory explained the dominance of word reading in terms of the relative speed of processing (e.g., Dyer, 1971). According to this account, in a kind of a horse race, faster word reading activates an incorrect response first. It takes time to deactivate this response, and to activate a response primed by a slower color naming process. Consequently, there will be little or no reverse effect, because the latter process is too slow to interfere with the former one. Another theory (Posner & Snyder, 1975) predicted that greater practice makes the dominant process independent of attention (difficult to control) and ballistic (obligatory). The less trained process is subject to attentional modulation and can easily be voluntarily withheld. Thus, the former process interferes with the latter, but not vice versa.

Glaser and Glaser (1982) questioned these two theories by presenting words up to a few hundred milliseconds before or after color patches (i.e., SOA from −400 ms to 400 ms; Simulation 11). They showed that interference dropped substantially when a word either preceded or followed a color by more than 100 ms. Only if both co-occurred, would typical interference arise. If relative speed mattered, then slowing down the color presentation should increase the interference, as the reading process would reach a respective response even sooner than would the naming process. In

contrast, slowing down the word presentation should result in the reverse Stroop effect (both processes would reach responses at the same time). Both of these outcomes were absent in Glaser and Glaser's data. Moreover, if the level of automaticity mattered, then it would be difficult to explain why a longer exposition to the prime that yielded the highly practiced process of word reading almost eliminated interference. Thus, the SOA effect seems to be a powerful test of existing Stroop theories and models (see Roelofs, 2003). In the late 1980s, in order to explain the SOA and other experimental effects observed in the Stroop task, the formal theories of human performance on this task were developed, usually expressed in the form of computational models capable of quantitative predictions. Early models primarily implemented the regulative function of cognitive control, that is, the modulation of the nondominant process (its activation, biasing, strengthening, prioritizing, etc.), increasing the likelihood of its successful application. More recent models also accounted for the evaluative function of control, responsible for judging the necessary involvement of the regulative function.

Stroop Models Based on the Regulative Function of Cognitive Control

With the advent of connectionist modeling (McClelland & Rumelhart 1986), several neural network models of the Stroop task were developed. These models shared an assumption that reading words and naming colors represented two distinct processing pathways that competed for output. The former pathway was more likely to win the competition, so an additional activation had to be passed from some task-demand nodes (representing attentional focus), in order for the latter pathway to win, but at the cost of a greater number of processing cycles (a larger response latency) and/or output errors.

In this line of research, Phaf, Van der Heijden, and Hudson (1990) proposed a model (called SLAM) that assumed the architectural difference between the word reading and color naming paths. Whereas the word input was linked to a respective vocal response node in a straightforward way (e.g., a node perceiving word "RED" directly activated a node for saying "RED"), the activation from color input (e.g., color "red") reached the respective node only via intermittent conceptual nodes. The SLAM model yielded congruency effects (Simulation 1), but was unable to correctly replicate the SOA effect (Simulation 11).

In the influential model proposed by Cohen, Dunbar, and McClelland (1990), both the word and color paths had the same network architecture (i.e., connected input-hidden-output layers), but differed in weights of connections between input and hidden, as well as hidden and output, nodes of each path. The word path was more strongly connected than was the color path, so the former path could pass more activation from a word to the respective output node (which accumulated this activation until a threshold was reached), than could the latter path. Only with an additional activation flowing from the task-demand nodes ("name colors") to the output node associated with a color could the correct response be made. However, overriding the incorrect response required extra processing.

The model's predictions diverted from observations in three crucial points. First, it predicted the normal distribution of response latencies, even though an asymmetric distribution, with the substantial tail of the longest responses, is commonly observed (Mewhort et al., 1992; Simulation 2). Second, when the strengths of both processing paths were equaled (via training), the model predicted no congruency effect, contrary to MacLeod and Dunbar (1988), who showed a substantial congruency effect even for symmetrical processes (Simulations 5 and 13). Third, neither this model nor its modifications (Cohen & Huston, 1994; Stafford & Gurney, 2007) correctly replicated the SOA effect (Simulation 11). Thus, despite its enormous popularity in the literature, the model seemed to reflect some serious theoretical flaws.

Some other connectionist models were more focused on the neurobiology of processing in the Stroop task, for example linking particular processes with precisely defined brain structures and/or functions (e.g., Dehaene, Kerszberg, & Changeux, 1998; Herd, Banich, & O'Reilly, 2006). These models accounted for the nature of particular Stroop effects to a lesser extent.

Another line of Stroop models, called hybrid, integrated symbolic (e.g., condition-action rules, called production rules) and subsymbolic (connectionist) principles. For instance, Roelofs' (2000, 2003) influential model of the Stroop task (Weaver++) contained distinct mechanisms for color/picture versus word naming, resulting from assumed differences in language production architecture. The model included three levels of linguistic representations: concepts, lemmas (syntactic representations), and word forms (phonological representations). Color perception, via related concepts, activated a corresponding lemma that had to be retrieved in order to select a proper word form. In contrast, a perceived word was able to directly activate the relevant lemma and/or word. Production rules determined actions made to retrieve concepts and lemmas. Color naming was achieved through an additional verification process, regulated by a color concept acting as a goal, and focusing the system on the rules responsible for processing perceived colors, but not words. Due to a shorter route for words led from perception to response generation, the congruency effect emerged. The model allowed Roelofs (2003) to successfully replicate data from 16 classic Stroop experiments, including correct simulation of the SOA effect (Simulation 11), the spatial disintegration effect (Simulation 12), the smaller word-word/color-color interference (Simulation 13), and the semantic gradient effect (Simulation 14). Two other hybrid models based on assumptions similar as Roelofs' model have been implemented in ACT-R cognitive architecture (Altmann & Davidson, 2001; Van Maanen & Van Rijn, 2007; Van Maanen et al., 2009).

Lovett (2001, 2005) proposed a substantially different ACT-R model of the Stroop. She introduced a subsymbolic parameter for each production rule, called rule utility. This parameter reflected the learning of a rule's successes and failures in attaining previous goals of the model, representing the sum of expected effectiveness (likelihood that the rule will fulfill the next goal) and expected efficiency (inverse of time in which the rule will attain the next goal). In resolving the competition between word reading and color naming rules, conceptualized by Lovett as the choice of the most utile strategy, her model used rule utility estimates (reflecting the response-outcome associations) instead of Cohen et al.'s (1990) path strength estimates (reflecting the stimulus–response associations). Both methods depended on the amount of practice/efficiency, but only the former also accounted for the rule's effectiveness from a more abstract perspective of agent's goals. Moreover, this model went one step beyond a purely regulative function

of control by proposing an additional check of whether or not a chosen production rule matched task instructions (for similar solutions, see also Altmann & Davidson, 2001; Juvina & Taatgen, 2009).

Stroop Models Including the Evaluative Control Function

The models of regulative control function have provided many insights into the Stroop task, but they did not tell why the human control system exerts control regulation. In order to explain this issue, Botvinick, Braver, Barch, Carter, and Cohen (2001) asked three (interrelated) crucial questions: When should control start to regulate the response choice? How strongly should it regulate such a choice? When should it be withdrawn? Answering these questions, Botvinick et al. (2001) proposed that the core function of cognitive control consists of the monitoring of conflicts occurring within the cognitive system, as well as responding to these conflicts. Cognitive control constantly evaluates the level of conflict among potential responses, and resolves that conflict by enhancing goal-relevant responses with strength proportional to the conflict level. If control eventually resolves the conflict, the control strength can be (gradually) decreased.

Specifically, Botvinick et al. (2001) supplemented the Cohen et al. (1990) network with a conflict monitoring node that computed for the output layer the Hopfield energy equaling $E = -\sum\sum w_{ij}\, a_i\, a_j$, where: a denoted activation of an output node, w_{ij} was a negative connection weight between nodes i and j (reflecting their competition), and summation was computed for all competing units. In the case of two responses, and assuming w to be -1, the conflict simply depended on $a_1\, a_2$. In congruent trials, only one output node was primed and gained a positive activation, while the other node's activation was close to zero, so value E was low. In incongruent trials, both nodes were highly active, so value E was high. This value was used to drive the activation of the task node representing color naming, thus in the incongruent trials control signals sent by this node were stronger compared with the congruent trials. By using the above mechanism, Botvinick et al. (2001) simulated several conflict adaptation and dynamics effects (e.g., the proportion-congruent and Gratton effect; Simulations 6 and 8) that were outside the scope of Cohen et al.'s (1990) original model (for successful simulations of other effects/tasks, see also Jones, Cho, Nystrom, Cohen, & Braver, 2002; Yeung, Botvinick, & Cohen, 2004; Yeung & Cohen, 2006; for a review, see Yeung, 2012).

More episodically driven accounts of how control is engaged in the Stroop-like tasks (usually, the flankers task) have been proposed on the basis of the reinforcement learning (RL) theory (see Sutton & Barto, 1998). For example, Holroyd and Coles (2002; Holroyd, Yeung, Coles, & Cohen, 2005) proposed that RL mechanisms use each action episode to learn predictions about the value (i.e., better or worse outcome) of each stimulus-response combination, by encoding what stimuli preceded a response, which response was made, and what the respective feedback was. By instant calculation of the difference between the value of the most probable response in a particular context (i.e., a response option that had been most highly activated) and some expected value (e.g., how well the response had been made so far in that context), the RL module could detect when negative differences occurred (i.e., when the response is going to be worse than usual), then reinforcing task-appropriate behaviors, while attenuating improper ones. If observed and expected values are permanently mismatched, the RL module could update its predictions. Due to that mechanism, cognitive control was able to block a response that was going to be wrong.

The RL models (Alexander & Brown, 2011; Holroyd et al., 2005) aptly predicted some EEG waves in the flanker task, like N2 and ERN (Simulations 9 and 10). However, no RL model has actually been used to simulate a sufficient set of behavioral effects in Stroop. Due to the relative simplicity of the RL models, as well as their strong focus on modeling the brain correlates of control, it is unlikely that these models are capable of such a comprehensive simulation. Moreover, Yeung and Nieuwenhuis (2009), using EEG and the flanker task, demonstrated that when the conflict monitoring model and the RL-based model yielded contrasting predictions, the data did not support the RL-based model. Thus, although the RL view presents an elegant integrative account of the brain indices of cognitive control, possibly explaining performance in some other tasks (but see Nieuwenhuis, Schweizer, Mars, Botvinick, & Hajcak, 2007), to date it does not seem to provide a plausible account of the Stroop-like phenomena.

Finally, integrating the conflict monitoring and RL mechanisms, Verguts and Notebaert (2008; for similar, but less elaborated proposals see also Blais, Robidoux, Risko, & Besner, 2007; Davelaar & Stevens, 2009) proposed a network in which the conflict level was used to specifically modulate the association strength between the task node and the stimulus and response nodes, instead of just globally increasing the activation of the task node. As a result, their model could learn (via conflict-modulated RL) which particular stimuli/responses yielded conflicts, applying increased control only toward them (i.e., replicating the item-specific proportion-congruent effect; Simulation 7). Hence, this model demonstrated that the conflict monitoring theory might benefit from accounting for some specialized RL mechanisms.

It needs to be noted that the conflict monitoring theory was recently criticized, and several alternative models were proposed that correctly replicated the control adaptation and dynamics effects in Stroop without positing any conflict monitoring processes. For instance, brain correlates of conflict have been claimed to reflect the time spent on task, with more time needed for incongruent than for congruent trials (Grinband et al., 2011; but see Yeung, Cohen, & Botvinick, 2011). The Gratton and proportion-congruent effects were modeled via dynamic reciprocal interactions between input and goal layers in a neural network (Scherbaum, Dshemuchadse, Ruge, & Goschke, 2012), learning contingencies between stimuli and responses (Schmidt, 2013), and Bayesian prediction of conflict to come (Jiang et al., 2014; Yu, Dayan, & Cohen, 2009). These models represent promising lines of research on cognitive control, which may extend our understanding of its evaluative function, however it seems that no alternative model thus far generated predictions that could not be handled by the conflict-monitoring theory (though some alternative models seem more parsimonious than is the original theory).

Summing up, so far three main lines of computational models of the Stroop phenomena have been proposed: the neural networks based on Cohen et al.'s (1990) architecture, Roelofs' (2000, 2003) language production system (Weaver++), and Lovett's (2001, 2005) utility-learning model implemented in cognitive architecture

ACT-R. Research have also demonstrated that models of the regulative control function (for reviews, see De Pisapia, Repovs, & Braver, 2008; Roelofs & Lamers, 2007) need to be supplemented with the evaluative function in order to explain when and how strongly the top-down influence of control is exerted and when it should be withdrawn (for reviews, see Alexander & Brown, 2011; Holroyd & Yeung, 2011; Yeung, 2012). In this vein, the conflict monitoring theory (Botvinick et al., 2001) has greatly extended the Cohen et al. (1990) network. Other lines of solutions to the control evaluation problem include the RL mechanisms (e.g., Holroyd et al., 2005), as well as Bayesian inference (e.g., Jiang et al., 2014), but so far, their application to the explanation of Stroop phenomena has been relatively limited.

Goals of the Study

Although existing Stroop models uncovered many cognitive mechanisms underpinning the Stroop, to date no such model was able to provide a comprehensive explanation for the complete set of crucial Stroop effects. Roelofs' Weaver++ and Lovett's utility-learning models handled the practice and stimulus-related effects (e.g., the color-color, temporal and spatial disintegration, and semantic gradient effects), which are difficult or even impossible to obtain for the neural networks. In contrast, the latter models replicated the effects of conflict dynamics and adaptation (including EEG waves), accounted for by neither Weaver++ nor the utility-learning model (as they lack conflict evaluation mechanisms). There are also some effects that have not, thus far, been correctly addressed by any model, including the shape of the RT distributions in Stroop (Mewhort et al., 1992) and the small but significant reverse interference (Blais & Besner, 2006, 2007). Our primary goal is to show that the complete set of the crucial Stroop phenomena can only be replicated if various regulative and evaluative mechanisms (e.g., production rules, reinforcement learning, conflict monitoring, control adaptation) become integrated into one cognitive control system.

Moreover, so far the conflict monitoring theory defined conflicts as direct incompatibility among the response options' momentary activation, which reflects the strength of stimulus-response association resulting from practice. As aptly noted by Lovett (2005), the conflict level primarily pertains not to the practice/strength of competing response options (reflected in their current activation), but to their utility for an agent's goals, resulting from the history of their associations with positive (or negative) outcomes. Definitely, activity and utility of options may be mutually incompatible. A highly active option (e.g., strongly compulsive behavior) may not be utile at all. If another highly active, but utile option competed with the former, existing conflict monitoring models would predict a maximum level of conflict, though rational analysis suggests that in fact conflict is relatively low, and the nonutile option should be directly rejected. In contrast, if an option were a very utile action that was not active enough (e.g., effective behavior not yet sufficiently trained), conflict monitoring models would indicate low conflict between this utile but weak option and the utile and active option, though, rationally, choosing between two potentially highly effective options (though differing in activity) may be a highly conflicting situation. Hence, we aim to show that some Stroop phenomena can be explained only if the conflict evaluation mechanisms use information about estimated utility of competing actions.

Finally, as noted by Alexander and Brown (2011), two different response options may fulfill the same eventual task/goal (e.g., using either the front or rear break stops the bike), so conflicts may be more validly assessed among tasks/goals (see also Holroyd & Yeung, 2011). Such a situation often occurs in real life, but so far has never been examined in the Stroop context. We intend to also address this issue.

In the remainder of the paper, we present a computational model of cognitive control in the Stroop task (henceforth called *the integrated utility-based model*), which is the production system that integrates (a) utility learning (Lovett, 2005) and outcome assessment (Alexander & Brown, 2011) applied to production rules (actions) with (b) the mechanism of rule conflict evaluation (Botvinick et al., 2001) that uses rule utility to regulate (c) the strength of exerted top-down control promoting the selection of goal-relevant over goal-irrelevant rules (Roelofs, 2003). Using the integrated model, we replicate the largest set (18) of crucial Stroop effects explained so far, including four effects addressed by none of the competing models.

The Integrated Utility-Based Model of the Stroop Task

We define the construct of cognitive control as a mental function which, for all currently available options (be it either actual responses, or possibly also cognitions like attention switch, memory retrieval, or taking a next reasoning step), *alters the probability distribution of option choices, from a distribution primarily determined by the learned effectiveness of options, into a distribution primarily determined by their goal-relevance* (an arbitrary distribution that depends on the current goal).

Model Architecture

The model relies on a hybrid, symbolic–subsymbolic production system, implemented in Common Lisp (its code can be downloaded from: http://ecfi-group.eu/index.php?s=downloads). Its cognitive processing results from three components: the *visual buffer*, the *goal buffer*, and a set of *production rules* (S-R rules). Specifically, the model can attend to one location [*X*, *Y*] on the virtual computer screen at a time, and can read into its visual buffer *chunk* V, representing one or more features (category, size, and/or color, etc.) of an object perceived. Chunk M, reflecting the *current goal* (e.g., task name), is maintained in the goal buffer. Crucially, each production rule is defined as a collection of conditions S (i.e., constraints imposed on the contents of the visual buffer) and a collection of actions R (i.e., commands to relocate the visual buffer, update the goal buffer, and/or initiate a response). For each rule, the *base utility* is defined ($U \in [0, 1]$), which reflects the rule's history of successes and failures in attaining goals. In a stochastic process, the conflict resolution mechanism selects one rule from all rules applicable in a given situation (i.e., the rules whose conditions match chunk V, which altogether constitute *the response set*), with a probability that is proportional to the *momentary utility* (U') of that rule, which is the function of its base utility as well as its *association* ($A \in [0, 1]$) with goal M (the less associated is a rule, the more its base utility is decreased). After applying the selected rule, the state of the system changes. A single

event of selection and application of a rule is called *a cycle*. With regard to these cognitive operations, our hybrid system is similar to other production architectures, especially ACT-R (Anderson & Lebiere, 1998) and Weaver++ (Roelofs, 2003), but is much simpler. Specifically, the system does not contain any declarative/semantic memory. It can also be used for modeling other cognitive control tasks, for instance the antisaccade task.

Our Stroop model also contains three control mechanisms. First, the model evaluates *the conflict level* ($C \in [0, 1]$) among the rules within the response set. Second, variable C is used to regulate *the control strength* ($G \in [0, g]$), where g is a free parameter defining the maximum possible control strength. These two processes are analogous to conflict monitoring and control regulation present in the Botvinick et al. (2001) model. Third, in a novel mechanism, for each rule, variable G is used in order to transform variable U to U'. The stronger the control strength (the larger G), the more the distribution of rule choice probability is determined by variables A (i.e., the goal relevance of rules), and the less it depends on values U (i.e., the learned effectiveness of rules). This mechanism directly implements our notion of cognitive control. These three control mechanisms are described in the remainder of this section. Figure 1 presents the overview of processing in the model, as well as the variables and parameters that are read/written at each stage.

First, before a new cycle starts, the model checks whether, in the preceding cycle, a goal has been attained. If yes, then in a simple reinforcement learning procedure (Sutton & Barto, 1998) the model adapts the base utilities of all rules that were executed between that goal and the latest goal preceding it:

$$U_{it} = U_{it-1} + \frac{f - U_{it-1}}{1 + L_i} \quad (1)$$

where U_{it} is the new value of the base utility of rule i in time t, f is the feedback value (ranging from zero, which reflects "complete failure", to one that indicates "full success"), and $L_i \in [0, \infty)$ is the number of cases in which reinforcement of rule i has been applied so far. In Stroop, variable f takes on a binary value describing whether the task instruction was fulfilled (e.g., the right key was pressed; $f = 1$) or not (a wrong key was pressed; $f = 0$), as perceived by a subject or signaled by the task, and L_i is simply the total number of trials responded to by rule i. The rationale for using L_i in Equation 1 is that the more evidence about utility value that

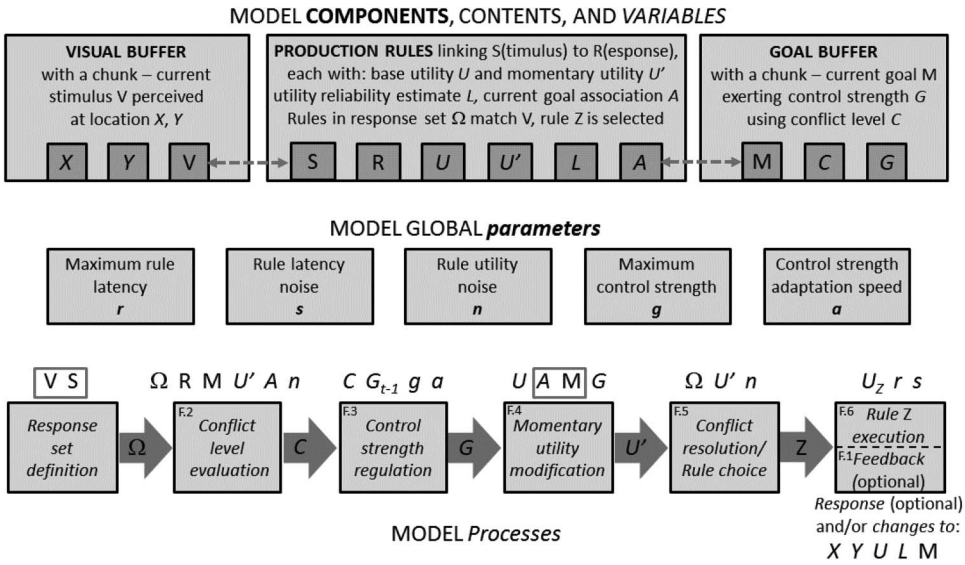

Figure 1. Architecture of the integrated utility-based model. Top: The model's components: visual buffer, goal buffer, and production rules, as well as their symbolic contents (regular capitals): visual chunk V, current goal M, and condition-action pairs S-R, respectively. The model's variables (italicized capitals) can be both read and written by the model. Local variables *U*/*U'* (base/momentary utility), *L* (base utility reliability), and *A* (rule-goal association) are defined for each production rule. Global variables represent the current location of the visual buffer (*X*, *Y*) as well as the current conflict level (*C*)/control strength (*G*) evaluated/exerted by the goal buffer. Initial values of variables act as parameters that become modified by the model during a task. Dashed arrows (and gray borders in bottom panel) indicate two intercomponent links: Conditions S of production rules are matched to visual chunk V, and rule-goal associations *A* are determined by current goal M. Middle: Five global parameters (bold italicized) that can only be read (but cannot written) by the model include rule utility noise *n* that regulates how deterministic/stochastic (low/large values) rule selection is, maximum control strength *g* that regulates the extent to which the model relies on values *A* in rule selection, control speed adaptation *a* that regulates how quickly control strength becomes adjusted to conflict, maximum latency *r* of rules with minimum utility (decreasing proportionally with increasing utility), and variance *s* of noise added to actual rule latency. Bottom: The sequence of processes run within a single cycle. Input contents/variables/parameters for a given process are shown above each box, and a respective output content/variable is displayed within an arrow. Symbols F.X in top-left corners denote the numbers of formula(s) implemented by each process. For a detailed description of each element of the model, see the Model Architecture section.

is gathered (the more reliable it is), the less altered it will be by the next feedback. When a rule has not previously been applied, then after its first execution, its utility reflects the exact value of the feedback. After many applications of the rule, its utility becomes very robust, and it changes little with new feedback. Obviously, this is a simplifying assumption, because in real RL, the feedback becomes discounted, and more recent feedback is weighted more than is less recent one (as the environment might have already changed and the latter feedback may be no longer valid). As we usually modeled one or 200 trials (except for Simulation 5) in stable conditions (the Stroop task sequence did not change), this simplification could not substantially affect the results. Overall, the base utility reflects the expected probability of achieving the next goal by means of the actions to which rule *i* (directly or indirectly) contributes.

Second, conditions S of each rule are tested against the chunk V from the visual buffer, and the matching rules are selected into response set Ω. If one rule is selected, then this very rule will be executed. If more alternative rules match, then the variable *C* on cycle *t* is calculated as the following ratio based on the Boltzmann distribution formula (from now on, symbol *e* stands for the base of the natural logarithm):

$$C_t = \frac{\sum_{k \in \Lambda \neq i} e^{U'_k / n}}{\sum_{j \in \Omega} e^{U'_j / n}} \quad (2)$$

The numerator of the ratio contains the momentary utilities of all production rules *k* (the *conflicting rules*; Λ) from the response set that yield different outcomes than the outcome yielded by rule *i* (the *most goal-relevant rule*) possessing the maximum association (A_{iM}) with the current goal M. Whether the outcomes of two rules are the same or are different is established by comparing their action sides R (e.g., whether they command to press the same or a different or no key). The denominator of the ratio includes momentary utilities of all rules within response set Ω. In other words, the conflict level depends on how utile in total the rules are whose outcomes are inconsistent with the outcome of the rule most strongly linked to goal M (see The Rationale for the Conflict Evaluation Formula section). Boltzmann's *temperature n* determines how sensitive formula (2) is to differences in momentary utilities. When *n* is close to zero, then ratio *U'*/*n* tends toward extremely large values for the relatively moderate values of *U'*, so values *U'* of the conflicting rules have to be really close to value *U'* of the most goal-relevant rule for a substantial conflict to appear. When *n* is large, and then ratio *U'*/*n* tends toward zero, each rule counts as one (i.e., e^0), regardless of its actual value *U'*, so *C* simply equals the proportion of the conflicting rules to all rules present in the response set.

Third, using the variant of the Botvinick et al. (2001) formula, the current strength of control (*G*) is set as a function of the maximum strength of control *g*, as well as the current level of conflict *C*, the latter adapted to by the control strength at a certain speed *a*:

$$G_t = agC_t + (1 - a)G_{t-1} \quad (3)$$

Parameter *a* takes on values from zero (the model never adapts to the current conflict, but exerts control of some constant strength) to one (the model adapts to the current conflict immediately).

Fourth, when the nonzero strength of control *G* is imposed, then the momentary utility *U'* of each rule *i* is calculated as a function of both its base utility *U* and its discrepancy with goal M, weighted by *G* (when $G = 0$, then *U'* simply equals *U*). Calculation of *U'* is governed by a newly introduced formula:

$$U'_i = \frac{U_i}{e^{G(1-A_{iM})}} \quad (4)$$

where $A_{iM} \in [0, 1]$ is the value of the association of rule *i* with (i.e., how relevant *i* is to) the current goal M (so, $1 - A_{iM}$ reflects *i*'s discrepancy with M). Particular values of *A* are theoretically fixed on the basis of analysis of a given task (see the next section). Specifically, A_{iM} equals zero if rule *i*, when fired, never leads to the attainment of goal M, A_{iM} equals one when *i* always leads to achieving that goal, and A_{iM} equals some intermediate value when *i* yields goal M only on some occasions (e.g., sometimes it fails to fulfill that goal, or it also leads to other goals). For the purpose of the current study, values *A* were defined as either zero (a rule does not match the task instructions: it does nothing or responds to a word when the goal is to name colors) or unity (a rule matches the instructions: it responds to a color when the goal is to name colors).

Having established the conflict level, control strength, and momentary utilities, the model uses the conflict resolution mechanism in order to select one rule Z out of response set Ω, and then executes that rule. The choice of rule Z depends on the stochastic procedure based on the nonlinear Luce (1959) ratio:

$$P_i = \frac{e^{U'_i / n}}{\sum_{j \subset \Omega} e^{U'_j / n}} \quad (5)$$

where P_i denotes the probability that rule *i* will be selected, and *j* reflects all rules within response set Ω. Similar to Equation 2, temperature *n* determines the sensitivity of the rule selection to the differences in momentary utilities between rules. When *n* is close to zero (the model is deterministic), then ratio *U'*/*n* tends toward extremely large values for the relatively moderate *U'* values, and the rule with the highest value *U'* will always be selected, even if its value *U'* is only slightly larger in comparison to the rule with the second-highest value *U'*. When *n* is large (the model is stochastic), then ratio *U'*/*n* tends toward zero, and each rule counts as one, having the same chance of being selected as any other rule from the response set, regardless of the actual values of *U'*. All simulations relied on a moderate value of temperature ($n = 0.2$).

The (non-negative) latency of each cycle *t* of the model operation is defined by Equation 6:

$$Lat_t = re^{-U} + noise(s) \quad (6)$$

where time *Lat* of the application of selected rule Z in cycle *t* is the function of parameter *r* reflecting the maximum possible latency of rule application in the model (820 ms for rules of zero utility), and the latency noise, drawn from the normal distribution with the mean equal to zero, and *SD* reflected in parameter *s*. Component e^{-U} decreases rule Z latency as a function of its utility, on the basis of a rational assumption that more utile rules fire faster than do less utile ones (analogous to how production strength affects latency in ACT-R; see Anderson & Lebiere, 1998).

Specific Chunks and Rules for the Stroop Task

In order to simulate the Stroop task, we introduced a simple set of chunks representing colors and words, as well as production rules that operated on those chunks. Below, we describe the standard four-stimulus/response variant of the model that was used in Simulations 1–16. Two other variants were introduced in Simulations 17 and 18, when different numbers of stimuli and responses were examined. For the sake of readability, below we refer to the manual color-word variant of the Stroop task, although the model is defined on an abstract level, and the same form of the model would apply to Stroop variants relying on other types of stimuli (e.g., the picture-word variant) or response modes (e.g., the vocal variant).

Chunks represented possible colors (red, blue, green, yellow) and words (RED, BLUE, GREEN, YELLOW). One color and one word feature could be accessed at a moment and for a duration determined by the algorithm controlling the stimulus presentation and response collection. There was also one goal chunk—"name-colors" (or, alternatively, "read-words" in Simulations 11 and 16). Eight specific production rules could be tested against the contents of the visual buffer. Four rules ("name-color-X") defined responses to the color features (e.g., "name-color-red" rule: if visual buffer = red then press key Z), whereas another four rules ("read-word-X") linked responses to the word features (e.g., "read-word-RED": if visual buffer = RED then press key Z). We assumed that these rules were compiled during training to such an extent that they did not require any memory retrieval (Taatgen & Lee, 2003). Of course, the rules linking reading words to button presses were not explicitly trained, but we assumed that such links were formed implicitly (in particular, such rules could yield correct responses in some congruent trials). Overall, the rules in the model can be treated as symbolic analogues of the stimulus-response paths in Cohen and colleagues' neural networks (Botvinick et al., 2001; Cohen et al., 1990).

Two additional rules were introduced on theoretical grounds. One rule ("other") represented all other possible cognitions, which were inappropriate in the Stroop task context (distractions, ruminations, etc.). This rule yielded no response, though it matched any conditions, and consumed some processing time. Another rule ("wait") was a strategic rule that could be applied when the system did not have enough evidence/confidence to make a response (see below). This rule made the model wait one cycle.

The magnitude of simulated congruency effects depended on the values of goal associations (*A*s), as well as the initial values of base utilities (*U*s) and utility reliabilities (*L*s) defined for each of the 10 models' rules. This does not mean that the model had 30 free parameters, because the set of parameters was highly constrained. First, because of the extremely skilled nature of reading, the initial base utilities of all four read-word rules were set theoretically at the maximum possible level (each $U = 1$). Also, as reading yielded a lot of feedback in the history of an agent, the initial reliability of the utility of read-word rules was defined as high (each $L = 1{,}000$). However, the use of these rules was prohibited by the task instruction, so in consequence their association to the goal was $A = 0$. In contrast, all four name-colors rules fully conformed to task instruction, their goal association was $A = 1$, so they constituted the most goal-relevant rules for this task. As they were less skilled, they had a substantially lower value of initial reliability (each $L = 50$), in addition to their utility being lower than the maximum utility possible (each $U = 0.79$). Rule "other," as not yielding any outcomes, had both a very low initial utility ($U = 0.1$) and no goal association ($A = 0$).

Finally, the base utility of rule "wait" was set to a large initial value of $U = 2.9$ (the only rule allowed a utility larger than unity) to reflect that waiting to respond in uncertain situations yields extraordinary high feedbacks of not committing—often fatal—errors. As rule "wait" was a general heuristic, thus it was not associated with any particular goal, and its parameter *A* equaled zero. The rationale for rule "wait" was that it was likely, due to its large base utility, to fire in situations when the conflict was high (two or more alternative rules had similar momentary utility), but the control strength was yet relatively low. It gave the model time to build up the control strength in order to boost the momentary utility of the most goal-relevant rule(s) and to make a more univocal decision. At the same time, the increasing control strength decreased the momentary utility of rule "wait," so this rule became less likely to fire. However, in speeded tasks the relying on that rule might lead to frequent errors of omission. Thus, in such tasks the model would quickly decrease its base utility due to its RL mechanism and would respond fast. As we modeled a relatively unspeeded variant of the Stroop task, we set that rule's utility high. Because rule "wait" reacted to conflicts, rather than elicited them, it was not taken into account in the conflict evaluation formula (2).

All 15 parameters of the model are summarized in Table 1. Nine of these parameters were optimized to maximize the model's fit, while the remaining six were set at the theoretically justified values. Additionally, in a few simulations where it is explicitly declared, some parameters were altered, but in each case, the rationale for these alterations was precisely explained. Finally, in order to match data from experiments in which overall RTs largely deviated from the mean RTs from Experiments 1 and 2 (due to the use of different task variants or task conditions), RTs in Simulations 6, 7, 8, 10, 12, 13, 14, 15, and 17 were scaled by a constant depending on the experiment modeled. Such a scaling changed only the overall RTs generated by the model, but not the qualitative patterns of the effects present in the simulated data.

The Operation of the Integrated Utility-Based Model of the Stroop Task

The operation of the model was very simple. In congruent trials, congruent features of the stimulus (e.g., color: blue and word: BLUE) were placed in the visual buffer. Exactly one name-color and one read-word rule matched these contents (i.e., "name-color-blue" and "read-word-BLUE"). Also, rule "other" was always included in the response set, because it matched any condition. Because both the name-color (the most goal-relevant rule) and read-word rule yielded the same correct key press, usually the correct response (either naming or reading) could be made in one or a few cycles. As only rule "other" yielded conflict *C*, its value and the resulting value of control strength *G* were low. However, even such a low value of *G* negatively affected values *U'* of goal-irrelevant rules, and made the name-color rule likely to be selected.

In contrast, in incongruent trials, features placed in the visual buffer differed (e.g., color: blue and word: RED). Thus, the rules

Table 1
Default Values of Parameters Used in the Integrated Utility-Based Model of the Color-Word Task

Parameter symbol	Parameter description	Value	Parameter type
U (Read-word-X)	Word reading initial base utility	1.0	Fixed (set theoretically)
U (Name-color-X)	Color naming initial base utility	.79	Free (optimized)
U (Other)	Task-unrelated thought initial base utility	.10	Fixed (set theoretically)
U (Wait)	Waiting rule initial utility	2.90	Free (optimized)
A (Read-word-X), A (Wait), A (Other)	The goal-irrelevant rules' association with the goal	.0	Fixed (set theoretically)
A (Name-color-X)	The most goal-relevant rules' association with the goal	1.0	Fixed (set theoretically)
L (Read-word-X)	Initial number of feedbacks received by word reading	1000	Fixed (set theoretically)
L (Name-color-X)	Initial number of feedbacks received by color naming	50	Free (optimized)
a	Conflict adaptation speed	.10	Free (optimized)
b	Hick's constant	.75	Free (optimized)
g	Control strength (mean)	28.0	Free (optimized)
n	Noise level (temperature)	.20	Free (optimized)
r	Maximum possible rule latency (for rules of zero utility)	.82s	Free (optimized)
$r_{perception/motor}$	Perception/motor rule latency	.15s	Fixed (set theoretically)
s	Standard deviation of rule's application latency noise	.13s	Free (optimized)

Note. In the word-reading task, as the goal was "read words," the values of parameters A(read-word-X) and A(name-color-X) were reversed (equaled 1.0 and .0, respectively).

matching the buffer contents yielded divergent outcomes (i.e., "name-color-blue" and "read-word-RED"). When the control strength was low, and momentary utilities were close to base utilities, then the read-word rule had a larger momentary utility than did the name-color rule. As both the read-word rule and rule "other" constituted the numerator of the conflict ratio, the conflict was relatively high. However, the read-word rule was less likely to be selected, because in such a case, rule "wait" had a much larger momentary utility. So, the repetition of the decision procedure was most likely. Meanwhile, the control strength was increased in response to large conflict. Due to formula (4), the momentary utilities of both the read-word rule and rule "wait" decreased (as none of them was associated with the goal), and in one of consecutive cycles, the name-color rule, whose utility was not decreased, might become selected. Compared with the choices in congruent trials, in incongruent trials responses took more time, as usually there were more cycles in one such trial. This led to visibly longer response latency for incongruent than for congruent trials. Figure 2 illustrates the operations of the model.

Rationale for the Conflict Evaluation Formula

The question of how to conceptualize conflict was asked often in 20th century psychology (e.g., Lewin, 1935). In particular, Berlyne (1960) attempted to define general criteria for the conflict metric. According to him (p. 332), the "degree-of-conflict" function C for N mutually incongruent response tendencies, each of a certain strength, should have the following properties: (a) C should be continuous, symmetric, and non-negative (with $C = 0$ for $N = 1$); (b) C should be maximal when strengths for all response tendencies are equal; (c) C should increase with an increasing N; and (d) C should increase when all strengths are multiplied by any value above unity. Berlyne proposed several formulae for conflict quantification that fulfilled the above criteria, though he left the list open. As noted above, Botvinick et al. (2001, p. 630) in their Stroop model used a formula that satisfied the Berlyne criteria—the Hopfield energy. In the Botvinick model, this formula quantified conflict as the *product* of activation (strength) of competing responses.

A different conception of conflict comes from Festinger (1957), who formalized the level of conflict (in his terminology, dissonance) between incongruent psychological entities (dissonances vs. consonances) in situations when some such entities were not mutually incongruent. According to Festinger, the level of conflict equals the *ratio*, instead of the Hopfield product, of dissonances, weighted by their importance, to the weighted sum of all dissonances and consonances that are meaningful for a given situation (Festinger & Carlsmith, 1959, p. 204). A person's motivation (which could be interpreted as the control strength) to counteract dissonance depends on the conflict level expressed in such a way.

As long as all applicable responses yield mutually incongruent outcomes, the Hopfield conflict is equivalent to the Festinger conflict (both rise and drop in the same way). However, when the Berlyne assumption, which indicates that all applicable responses must be mutually incongruent, is broken, then the Hopfield formula no longer provides correct evaluation of the conflict level. Specifically, it predicts the level of conflict to decrease when additional dissonant response tendencies are introduced to the set of applicable responses, although, analyzed rationally, the conflict level in fact increases, as the chance of selecting a dissonant (unwanted) response rises. The above decrease in conflict level will be exactly the same as when the number of consonant responses increases, although, obviously, the latter case yields less conflict (a wanted outcome becomes more likely). In the same circumstances, the Festinger formula behaves correctly: More dissonant responses in the set of applicable responses yield more conflict, whereas more consonant responses yield less conflict. This line of reasoning is illustrated in Figure 3.

For this reason, we assume that the Festinger approach (related to the ratio of incompatible rules to all rules) is a much more appropriate measure of conflict than is the Hopfield approach (related to the product of rules weighted by their incompatibility) adopted by Botvinick et al. (2001). Simply, only the Festinger approach can easily generalize onto the situations that often occur in real life, that is, when two or more response tendencies pertain to a certain goal, even if they achieve this goal in different ways.

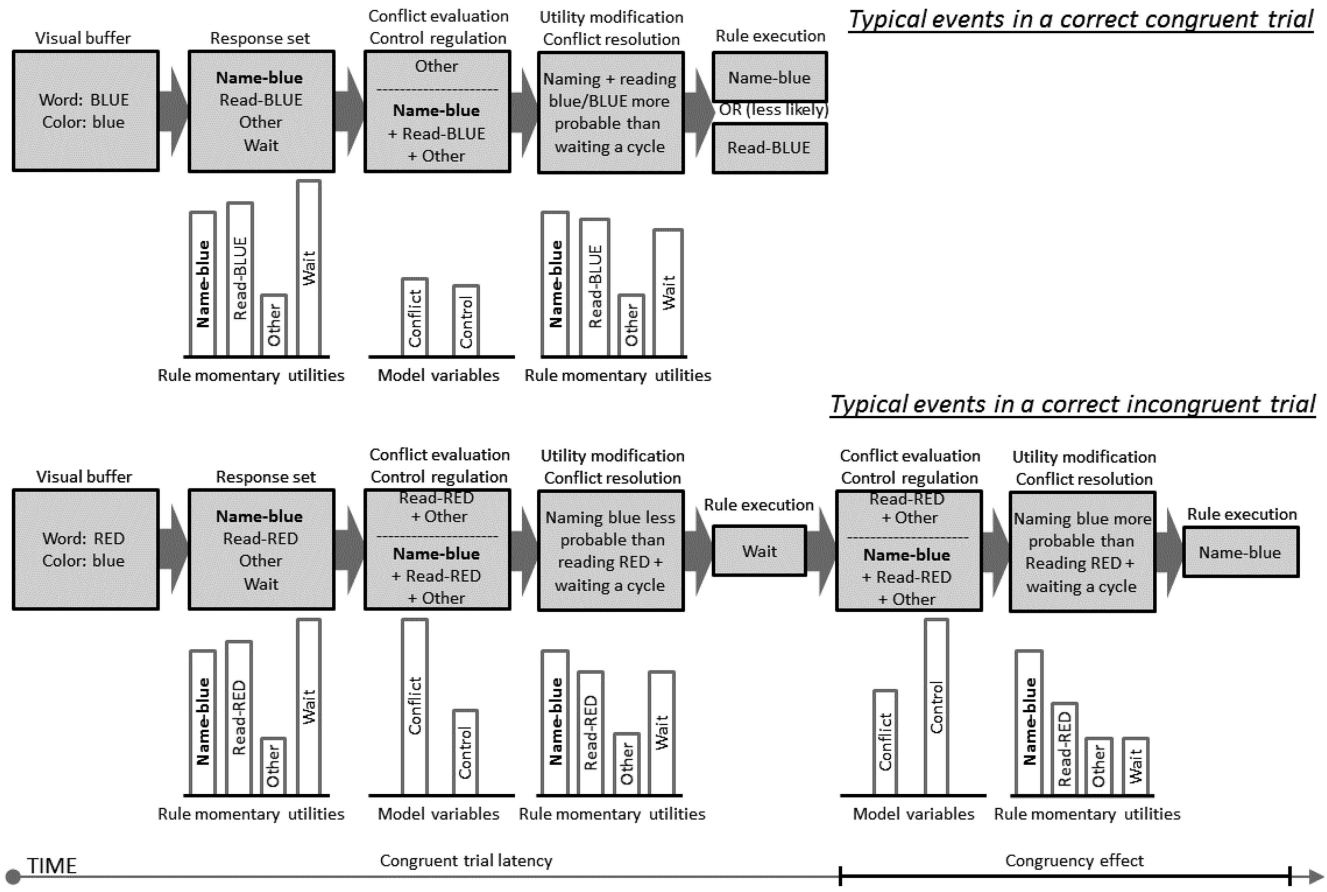

Figure 2. Typical events in a correct simulated congruent (top) and incongruent trial (bottom) of the Stroop task. In the congruent trial, the mutually matching color and its name are placed in the visual buffer. Four rules that: (a) name or (b) read this color (yield the same outcome); (c) represent "other," task-unrelated, process; and (d) strategically postpone response ("wait") all compete for selection from the response set. Each rule's initial momentary utility (*U'*), proportional to its chance for selection, is close to its base utility (*U*), proportional to the previous generalized effectiveness of the rule. The bold font denotes the most goal-relevant rule (= "name-color"), whereas the remaining rules are goal irrelevant. Then, conflict level *C* is evaluated proportionally to utilities of rules that yield different outcomes than the most-goal relevant rule. In the congruent trial, only rule "other" yields such an outcome (the strategic rule "wait" is not counted), so value *C*, as well as control strength *G* proportional to *C*, are low. However, even such a low value of *G* negatively affects values *U'* of goal-irrelevant rules, and makes the name-color rule most likely to be selected. As the summary probability of selecting a correct response (either naming or reading blue) is higher than is the chance of waiting one cycle, the model likely selects a response, whose latency is low. In the incongruent trial, color name and word activate two rules that yield different outcomes, so conflict *C* is proportional to the sum of the utilities of rules "read-word" and "other," both yielding different outcomes than the outcome of "name-color," and *C* is large. As control strength *G* adapts to conflict *C* with a delay, values *U'* of goal-irrelevant rules become decreased only little (similarly to the congruent trial), and the chance of naming a color is lower than is the summary probability of reading a word (leading to an error) or waiting one cycle. Thus, the model usually needs more cycles to increase the relative probability of color naming, and selecting the correct response. Latency of these additional cycles constitutes the congruency effect widely observed in the Stroop task. For more information, see the Specific Chunks and Rules for the Stroop Task section and The Operation of the Integrated Utility-Based Model of the Stroop Task section.

The present model implements the Festinger approach in the form of Equation 2. In Simulation 18, we show the specific consequence of the Festinger formula that consists of an increase in evaluated conflict (as measured with EEG) when relatively more potential responses are linked to an incongruent (i.e., word meaning) aspect of a stimulus (i.e., a colored word).

Simulations

First, two experiments presented in Appendix were conducted in order to collect data for the basic model fits. Experiment 1 was applied in order to collect a large dataset on the individual differences in congruency effect in a sample of 340 participants in the

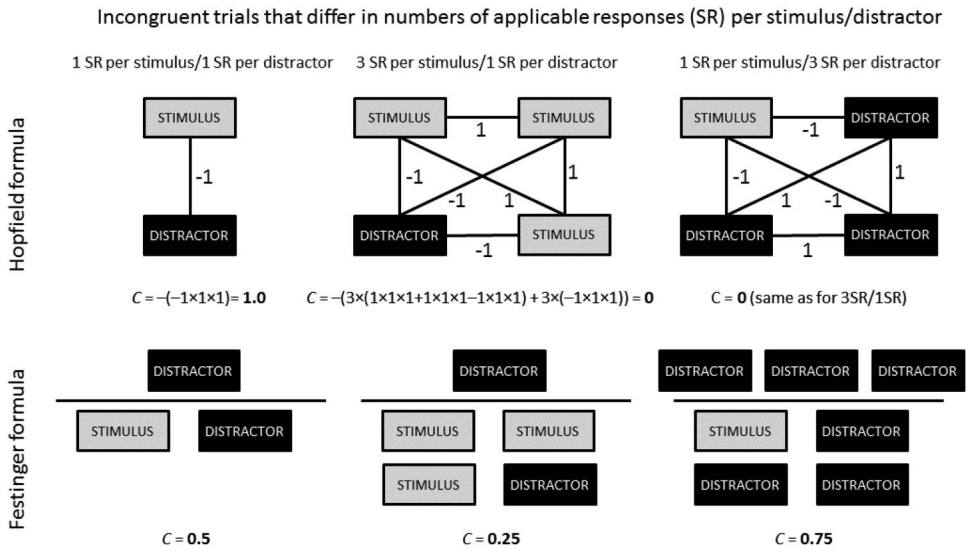

Figure 3. Illustration of the conflict level (C) evaluation according to the Hopfield versus Festinger formula. Left: An incongruent trial in the standard Stroop in which there is one applicable manual response per target color and one per distractor word. Middle: Analogous trial in the modified Stroop in which three different responses (e.g., three buttons associated with three fingers) can be correctly applied to a color, but only one response can be (incorrectly) applied to a word. Right: Analogous trial in which one response can be correctly applied to a color, but three different responses can be (incorrectly) applied to a word. Numbers by the lines represent connection weights in the variants of the Botvinick et al. (2001) model that directly implemented the Hopfield formula.

color-word and picture-word tasks, which were needed for the distributional analyses of RTs, as well as for the Stroop task variants comparison (Simulations 1–4). Experiment 2 examined the adaptation effects in Stroop (Simulation 6), escaping some problems of the previous studies that were pointed out by Melara and Algom (2003). In most simulations, each of 340 simulated agents performed 100 trials of the Stroop task, matching the sample size and trial number for Experiment 1. For each agent, the value of maximum control strength (parameter g) was randomly drawn from the exponential distribution in order to reflect individual differences in executive control commonly observed in the population. For all simulations, we present 95% confidence intervals, indicating the respective variance in simulated data. A total of 18 simulations were divided into four thematic sections: basic congruency effects (Simulations 1–5), performance dynamics and adaptation (Simulations 6–10), experimental manipulations made to stimulation (Simulations 11–14), and manipulations to responding (Simulations 15–18). Although we modeled effects selected from a larger pool of existing Stroop phenomena, on the basis of MacLeod (1991) and Roelofs's (2003) reviews we focused on the most important effects (i.e., most widely examined), and in particular, we chose effects that most clearly differentiated the existing models in how they account for the Stroop (i.e., at least one model did not replicate a given effect). For discussion of other effects, see the Comparison with alternative models of Stroop section.

Basic Congruency Effects

Simulation 1 (interference exceeds facilitation). In Simulation 1, we replicated the congruency effects found in Experiment 1. All existing models of Stroop successfully replicated such effects. The present model generated the correct error rates for both the congruent (M_{sim} = .99 vs. M_{obs} = .97) and incongruent trials (M_{sim} = .92 vs. M_{obs} = .90) of the color-word task, yielding the congruency effect of M_{sim} = .076 (M_{obs} = .069). Also, the correct latency effect of M_{sim} = 129 ms was obtained (identical as M_{obs} = 129 ms; for the underlying distributions see Figure 5).

The congruency effect consists of interference and facilitation, and the latter is usually smaller than the former (Dyer, 1971; Stroop, 1935). Two influential explanations of the facilitation effect exist. One (Cohen et al., 1990; see also Roelofs, 2003) assumes that when a color and a word lead to the same response (in congruent trials) the summation of activation/strength of the word reading and color naming response paths makes people respond faster, compared with neutral trials, when only one path—related to color naming—can be elicited. The alternative explanation (MacLeod, 1991) predicts the speed increase in the congruent trials in terms of frequent fast responses based on word reading, which cannot occur in neutral trials. Both of these plausible explanations were supported by existing empirical data. So far, only the Lovett (2005) model integrated both of these explanations in one facilitation effect.

Our model also integrates these two sources of facilitation. First, it predicts that in congruent trials, the denominator of the conflict Equation 2 is larger (because it includes the utilities of both the name-color and read-word rules) compared with neutral trials (only the utility of the name-color rule), thus the conflict is lower in the former trials. Second, the model predicts that in a certain proportion of congruent trials, it is the faster read-word rule that generates a response. Figure 4 compares latencies in incongruent,

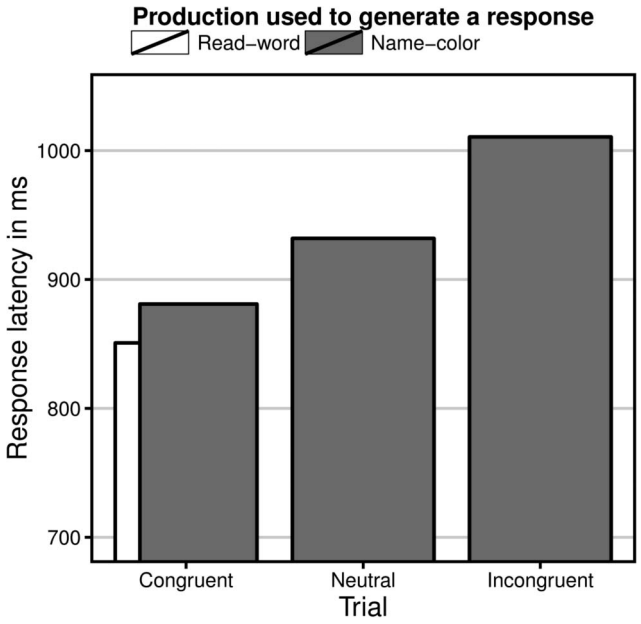

Figure 4. Mean response latency in the simulated incongruent, neutral, and congruent trials, for the latter case separated according to whether a response was based on either the read-word or the name-color rule. The relative width of the bars represents the proportion of responses yielded by a given rule.

neutral, and congruent trials, the latter divided according to whether the name-color (83% trials) or read-word rule (17% trials) generated a response. The model replicated a visibly smaller facilitation effect (56 ms) compared with the interference effect (79 ms), and the former effect resulted from both the Cohen et al. (1990) convergence of path strengths and the MacLeod (1991) occasional reading of words in the congruent condition.

Simulation 2 (congruent/incongruent trials latency distribution). For any Stroop model, it is crucial to show not only that it correctly replicates mean response latencies in Stroop, but also generates the RT distributions that yield those means in a similar way as it is observed for people (see Mewhort, Brown, & Heathcote, 1992). Top panel of Figure 5 presents, using all data points from Simulation 1, the kernel density estimations of latency distributions in congruent (dashed lines) versus incongruent trials (solid lines), matching the shapes of respective distributions observed in the color-word task from Experiment 1. First, both distributions approximated the ex-Gaussian distribution (not the normal one). Second, the incongruent distribution was more platykurtic than was the congruent one. Third, the former displayed a larger tail than did the latter. Bottom panel of Figure 5 presents the mean values of the .1, .3, .5, .7, and .9 RT quantiles, averaged over the individual quantiles of each of 340 participants/agents.

Furthermore, we used the ex-Gaussian theoretical distribution in order to precisely model both the observed and simulated distributions. The ex-Gaussian function represents the convolution of the normal distribution with mean μ and standard deviation (*SD*) σ, and the exponential distribution with mean (as well as *SD*) equal to τ (i.e., the inverse of λ). The mean of the resulting distribution is equal to $\mu + \tau$. Component μ reflects the central tendency of the leading edge of distribution, whereas component τ represents the magnitude of its tail. It is widely held that real RT distributions, due to their commonly visible right-side tails, can be better described by the ex-Gaussian (and other asymmetric distributions, like shifted Wald, shifted Weibull, and shifted log-normal) than by the sheer Gaussian function (e.g., Ratcliff, 1979). Although to date

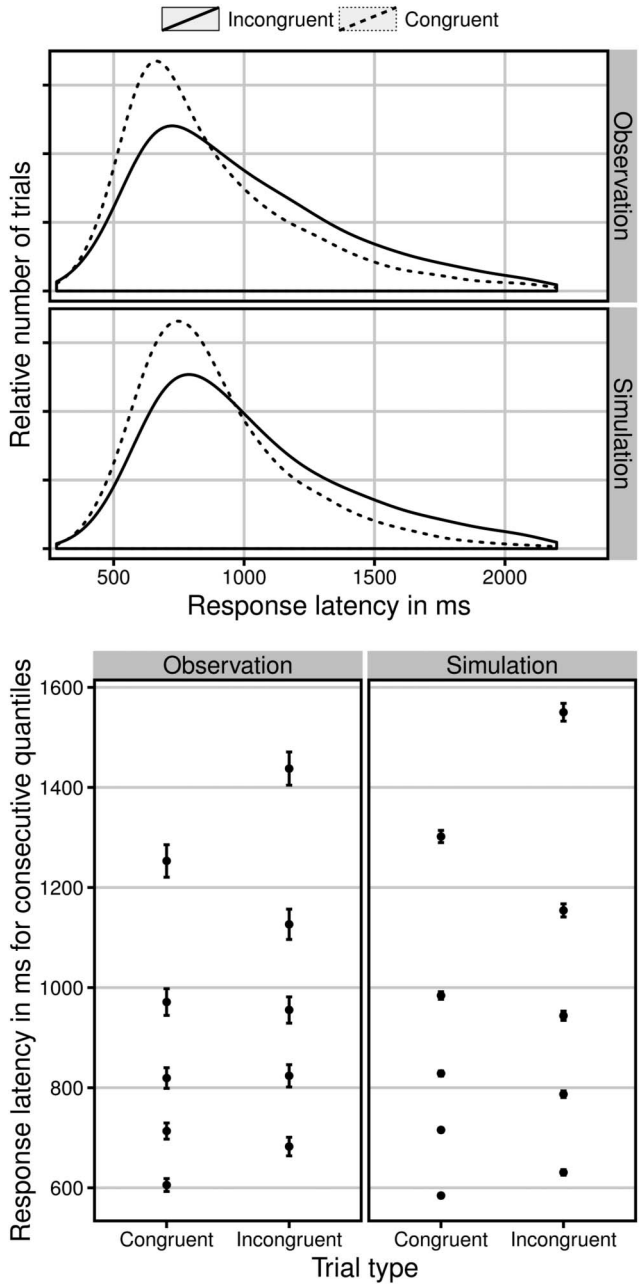

Figure 5. Top panel: Kernel density approximation of the distributions of observed (Experiment 1) and simulated response latency in congruent and incongruent trials. Bottom panel: Quantile probability plot for response latency. Each point represents mean of individual quantile values for .1, .3, .5, .7, and .9 quantiles, from bottom to top, respectively. Error bars represent 95% confidence intervals for means.

mechanisms causing differences in both μ and τ in various conditions of the Stroop task have been actively debated (for opposing views, see Roelofs, 2012; Spieler, Balota, & Faust, 2000), ex-Gaussian modeling was helpful in testing the Stroop models. In one such study, Mewhort et al. (1992) showed that although Cohen et al.'s (1990) model validly tapped changes in mean RT between neutral, congruent, and incongruent conditions of the Stroop task, it failed to replicate the exact changes in μ and τ. Specifically, depending on the model parameters, in incongruent trials, it predicted an increase in either μ or τ, but not both at the same time. The Roelofs (2003) and the Lovett (2005) models have never been used to generate latency distribution in the Stroop. We fitted the ex-Gaussian parameters with the Markov chain Monte Carlo algorithm (slice sampling; Neal, 2003) to both the observed and the simulated data. The estimated parameters of each RT distribution closely matched (see Figure 6). Moreover, in both the model and the observations, the increase in mean RT in the incongruent condition, compared with the congruent one, resulted from both an increase in parameter μ ($\Delta\mu_{sim}$ = 23 ms and $\Delta\mu_{obs}$ = 61 ms) and parameter τ ($\Delta\tau_{sim}$ = 98 ms and $\Delta\tau_{obs}$ = 63 ms). Thus, a larger mean latency in the incongruent trials resulted from both a general slowing of responses and a larger tail of the longest responses.

The close similarity between the simulated and observed distributions was also shown by the Kullback-Leibler divergence statistics equaling extremely low values of 0.032 for the congruent condition and 0.012 for the incongruent one. This statistic is a logarithmic, weighted difference between normalized areas of continuous probability distributions, reflecting the proportion of the area of two distributions that differs between them. Here, only 2% of the area differed on average. Thus, our model was the only one that correctly accounted for the differences in the observed response latency distributions between congruent and incongruent trials (and not only predicted the mean RT difference).

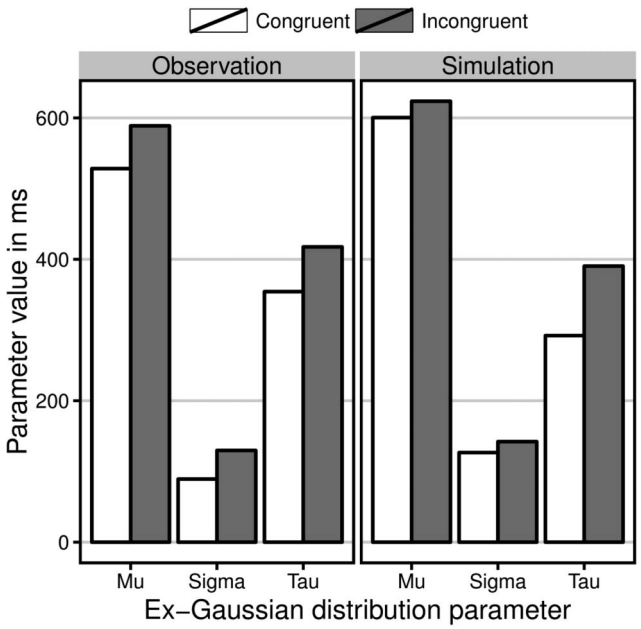

Figure 6. Ex-Gaussian distribution parameters in observed (Experiment 1) and simulated data.

Simulation 3 (congruency effect distribution). The mean congruency effect, observed in Experiment 1, obviously resulted from some distribution of mean congruency effects displayed by individuals. As we examined a sufficiently large sample, we were able to analyze that distribution, shown by dashed line in top panel of Figure 7. Participants displayed large variation in how they coped with conflict, with effects ranging from insignificant to substantial. As far as we know, no existing model of Stroop has attempted to replicate the pattern of individual differences in congruency effect, although each group of models includes a parameter, like the goal-node weight (neural networks), the total amount of activation spread from the goal (ACT-R's parameter *W* in the utility-learning model), or the control over distraction (parameter *du* in Weaver++) that, when manipulated, could potentially yield a certain distribution of congruency effects, whose shape likely depends on the parameter values applied.

How can our model account for such differences in Stroop performance? An analysis of Equations 1–6 suggested that five parameters might affect the time in which the model responds to incongruent trials. First, the higher the value of parameter *g* that regulates the momentary strength of control, the more *U'* of the read-word rule decreases, and the resulting conflict (as well as the congruency effect) is lower. Second, parameter *g* may not vary, but the faster the model adapts the strength of control to the level of conflict (i.e., due to parameter *a*), the stronger the control exerted on average, possibly leading to similar effects to the effects yielded due to variation in parameter *g*. Third, the level of conflict may somehow be influenced by noise parameters *n* and *s*. Fourth, an increase in base utility *U* of the name-color rules may result in their winning over rule "wait" in a lower number of cycles, thus, in a shorter RT.

The latter option was discarded, as it was unlikely that young adult people from Experiment 1 differed substantially in the base utility of the name-color rule, as these participants had a relatively similar educational history (see Simulation 13 for parametrical analyses pertaining to base utility). For the four remaining parameters it was tested whether manipulating their values could generate the observed latency congruency effect distribution. The results indicated that only manipulating parameter *g* yielded the correct variation in this effect, whereas it was barely sensitive to changes in parameters *a*, *n* and *s*. Consequently, values *g* were randomly picked from the exponential distribution with $M = 28$ and minimum = 4. Solid line in top panel of Figure 7 shows the replicated distribution. Moreover, simulated data (middle panel) nicely matched the distribution of accuracy in the incongruent trials (the congruent trials always yielded close-to-perfect accuracy, so we ignore them). The model also captured the negative correlation between latency effect and accuracy ($r = -.49$; $p < .001$), which was present in the observed data ($r = -.30$; $p < .001$). Finally, bottom panels of Figure 7 present individual RT distributions for four randomly picked agents that differed in parameter *g*, as well as four example human participants ordered according to the decreasing congruency effect. The analysis of the influence of parameter *g* on the congruency effect and accuracy is shown in Figure 8. Overall, the simulated and observed distributions closely matched, as shown by the very low values of the Kullback-Leibler

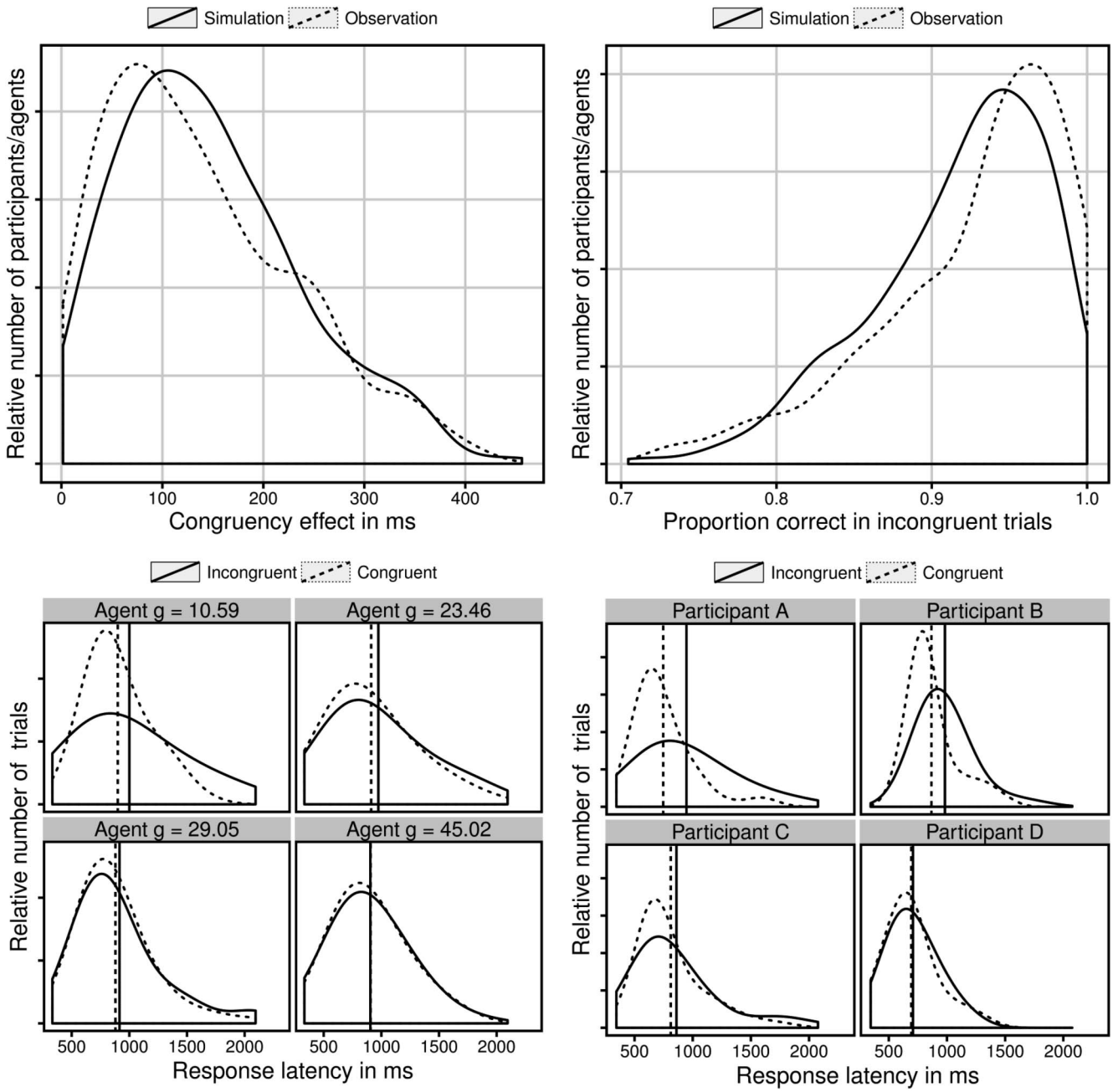

Figure 7. Top panels: Kernel density approximation of the observed (Experiment 1) and simulated distributions of the latency congruency effect (left) and the incongruent trials' accuracy (right). Bottom panels: Distributions of RT in four example participants ordered by the decreasing congruency effect (left), as well as RT in four example simulated agents ordered by the increasing maximum control strength g. Solid/dashed vertical lines denote mean latency in the in/congruent trials. The difference between the lines represents the congruency effect, which in the agents visibly decreases with their increasing values of g.

divergence statistics equaling 0.018 (latency congruency effect) and 0.033 (accuracy).

Simulation 4 (smaller congruency effect in the picture-word task). In his review, MacLeod (1991, p. 170) concluded that diverse Stroop variants yielded similar patterns of data, suggesting that they may be undergirded by the same cognitive mechanism. In Experiment 1 we found that the figure-word effect was much smaller than was the color-word effect. So, assuming that the same cognitive mechanism yields both effects, what may vary between the two tasks? Each of the existing Stroop models attempted to account for the picture-word effect. The Cohen et al. (1990, see their Figure 10) and Roelofs (2003, see his Figure 16) models

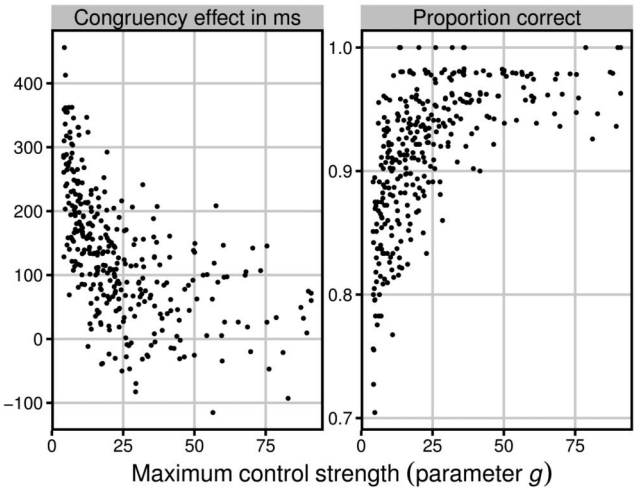

Figure 8. Relationship between parameter *g* (the maximum control strength) and the simulated size of congruency effect (left) as well as the proportion correct in the incongruent trials (right).

assume that exactly the same mechanism is recruited to cope with the interference from word reading when naming both colors and pictures. Van Maanen et al. (2009) proposed a more specific account. Although they also assumed that generally the same mechanism is responsible for color and picture naming, they additionally varied a parameter reflecting the speed of perceptual processing of colors (faster processing) versus pictures (slower processing). This allowed them to capture a difference in the locus of interference in both tasks, with the picture-word interference occurring primarily during the perceptual stage, whereas the color-word interference pertained mainly to response selection. However, their model incorrectly predicted much larger facilitation than interference in the picture-word task (see their Figure 1)—a pattern never observed in real data (e.g., Dyer, 1971; Glaser & Glaser, 1982; Tzelgov et al., 1992). This fact leaves room for other explanations of a possible difference between the color- and picture-word interference, beyond sheer perceptual processing time.

Consequently, we assumed that the utility of the nondominant process (i.e., naming) varies between color- and picture-word tasks (note that reading is the same for both tasks). In Simulation 4, we tested this prediction by setting a larger utility for the figure than for the color naming, inferred from the overall lower RTs for the figure-word task. We resimulated each agent from the color-word task simulation, now with the utility of the name-figure rules set to 0.90 (in comparison with 0.79 set for the name-color rules). Figure 9 shows data for both variants of the Stroop task. It can easily be seen that the simulated congruency effect in the figure-word task decreased by half in comparison with the color-word task, matching the observed data. Moreover, unlike in van Maanen et al. (2009), the interference component of the figure-word congruency effect was visibly larger than was the facilitation component.

Simulation 5 (practice reverses interference). MacLeod and Dunbar (1988) showed that the intensive training of the nondominant process leads to the reversal of interference. After training, this process starts to interfere with the originally dominant process, but little interference is observed vice versa. As training the color naming to be better learned compared with the word reading is impossible (MacLeod, 1998), MacLeod and Dunbar designed a Stroop task in which, first, people learned arbitrary associations between colors and uncolored shapes (the training phase), and then they provided associated names for the colored shapes (the test phase). In the case of incongruency between the to-be-named associated and the to-be-ignored actual color of a shape, the same congruency effect showed up as in the classic Stroop, whereas when the task required naming actual colors, the learned color-shape associations barely interfered. Then, after 5 days of the training procedure, the associated-color naming task sped up, and conflict between the associated and actual colors started to yield a similar congruency effect in both tasks. Finally, after 20 days of training, the speed of naming an associated color surpassed that of naming an actual color, and the interference pattern reversed. Now, actual-color naming yielded a congruency effect, whereas the associated-color naming did only a little.

Cohen et al. (1990) modeled this effect by teaching the network the relevant associations between stimuli and responses for the shape-naming path. As noted, their model failed to correctly predict the MacLeod and Dunbar pattern of data, because it yielded no congruency effect if both tasks were equally learned (i.e., after a 5-day training), whereas MacLeod and Dunbar observed the symmetrical interference both for actual and associated color naming. In contrast, two other models correctly replicated the mutual interference on the fifth day. The Roelofs (2003) model accounted for learning effects by strengthening the associations of a shape with a concept and lemma associated with that shape. In the Lovett (2001, 2005) model, the gradual increase in the utility of a rule responsible for naming associated colors yielded an eventual advantage of this naming process over the naming of actual colors.

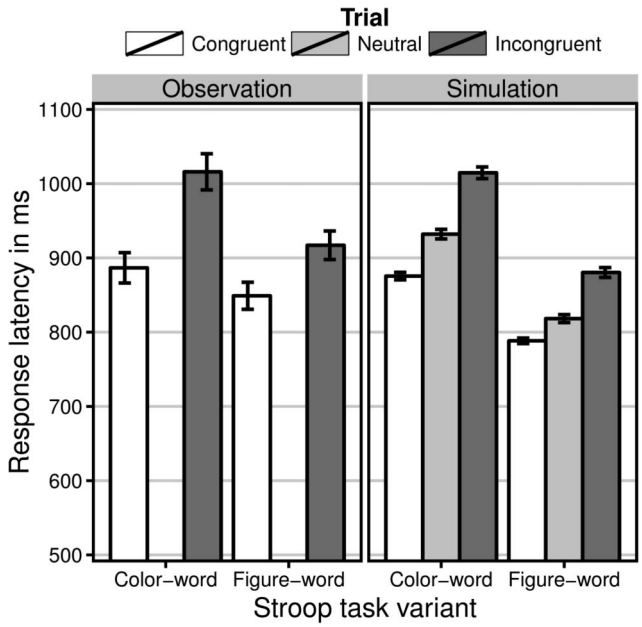

Figure 9. Mean response latency in the congruent and incongruent trials in observed (Experiment 1) and simulated data for the color-word and figure-word variants of the Stroop task. The neutral trials were also simulated for each variant in order to demonstrate visibly smaller facilitation than interference. Error bars represent 95% confidence intervals.

Similar to the Lovett model, the present model also used its RL mechanism to increase the utility of associated color naming throughout the learning phase. The model was tested for the associated and actual-color naming tasks after 0, 2,664, and 10,656 trials (similar to the original study) of naming colors associated with uncolored shapes. Three parameters were fitted: the utility of shape color naming ($U = 0.75$), the utility of actual color naming ($U = 0.90$), and the latter rule's reliability ($L = 1000$). Figure 10 shows that the simulated data nicely matched the MacLeod and Dunbar data. At the start of training, there was a larger interference for naming shape, compared with actual, colors; in the middle, the interference for both processes was identical; but at the end, the interference for naming actual colors surpassed the one for associated colors. Overall, the simulated effect resulted from the fact that the RL formula (1) yielded a permanent increase in the utility of shape color naming, at some point surpassing the utility of actual color naming. Bottom panel of Figure 10 shows the changes in utility of both rules.

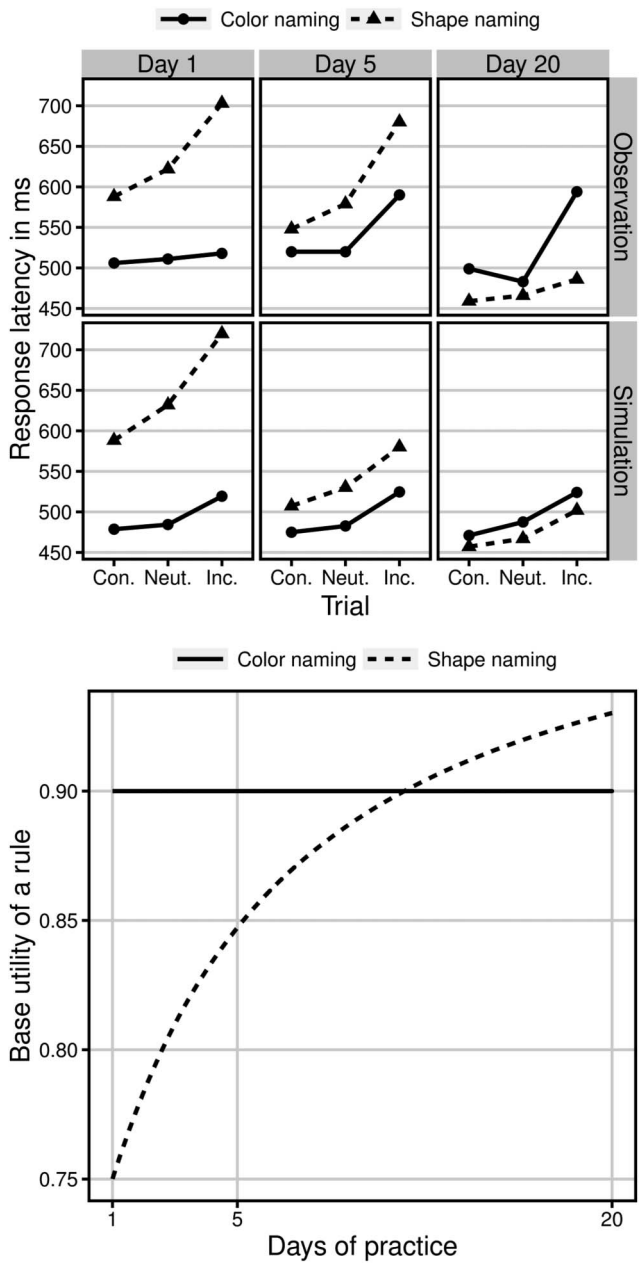

Figure 10. Top panel: Observed (MacLeod & Dunbar, 1988, Experiment 3) and simulated mean response latency in the congruent, neutral, and incongruent trials of naming colors associated with shapes versus their actual colors, at the start (Day 1), middle (Day 5), and end of training (Day 20). Bottom panel: Practice-driven changes in the base utility of respective production rules.

Dynamics and Adaptation of Control

In this section, we simulated the effects that pertained to the dynamic changes in conflict and/or control resulting from adaptation to the (in)congruence of trials and/or trial sequences. These effects (i.e., the proportion-congruent and item-specific proportion-congruent effects, the preceding trial's [Gratton] effect, and the EEG indices of control) were the main arguments in support of neural networks (Botvinick et al., 2001; Jones et al., 2002; Verguts & Notebaert, 2008; Yeung et al., 2004), which, owing to their conflict evaluation mechanism, easily replicated all effects discussed in this section. In contrast, the Roelofs (2003) and Lovett (2005) models, simply because they lack any conflict evaluation mechanisms, so far have not attempted to replicate those effects (except for the proportion-congruent effect accounted for by Lovett, 2005), and in principle, it is difficult to explain how they could replicate them (especially EEG data, as no counterparts for the brain correlates of conflict/control have been defined for those models).

Simulation 6 (proportion-congruent effect). A larger interference in the primary congruent (PCS) than in the primary incongruent sequences (PIS), observed in Experiment 2, was successfully generated and is depicted in top panel of Figure 11.

The dynamics of conflict evaluation (variable *C*) and the resulting control strength (variable *G*), presented in bottom panel of Figure 11, clearly show that the proportion-congruent effect is rooted in the model's permanently higher conflict/control in PIS that results from the model's inertia in adaptation of the control strength to the conflict level. A permanently increased control strength in PIS attenuated the dominance of the read-word rule, as its value *U'* was constantly lower than in PCS, whereas *U'* of the name-color rule remained unchanged. As a result, PIS yielded smaller congruency effects. Because in the present model, the relative difference between momentary utilities matters for the conflict level (see Simulation 13), the decrease in interference in PIS can equally well be interpreted as the attenuation of word reading (see Lindsay & Jacoby, 1994) and the strengthening of color naming (see Botvinick et al., 2001).

The solution to the proportion-congruent effect adopted here is conceptually similar to one present in the Botvinick et al. model, which also easily replicates that effect. Lovett (2005) proposed another solution, by using more frequent reinforcement learning of the name-color rules in PIS that boosted their utility (and also speed), resulting in less interference in PIS than in PCS (though Lovett described only a general mechanism yielding the effect, but did not report the exact simulation data). In the next simulation, we show how RL can handle the item-specific variant of the

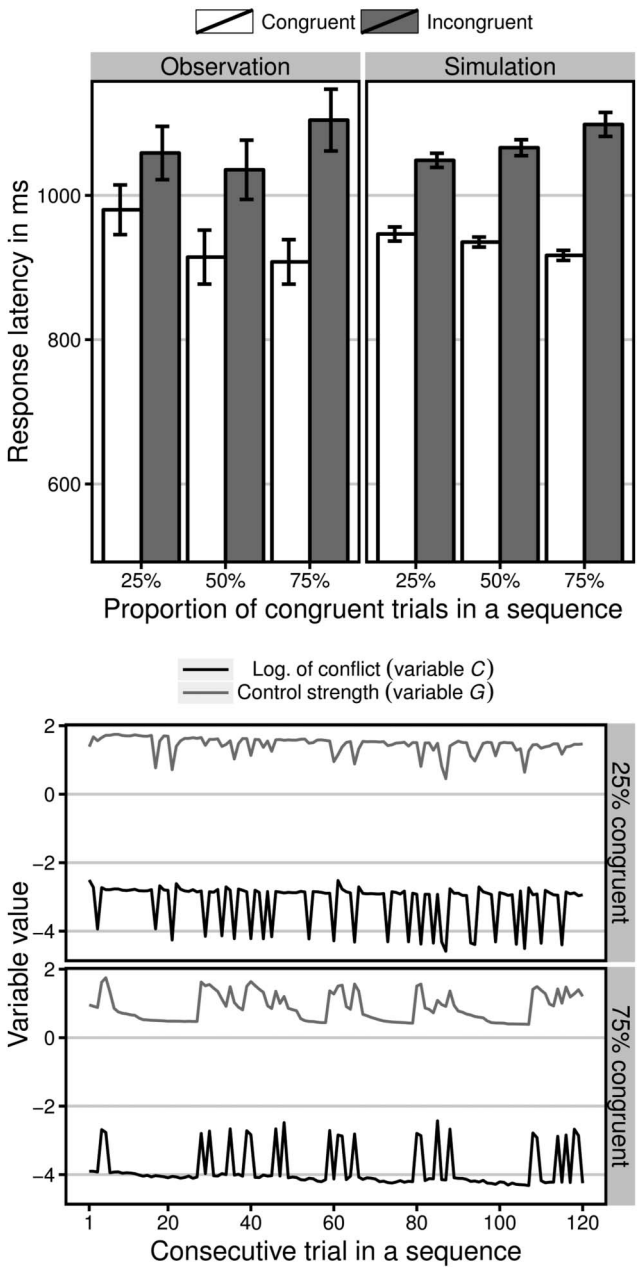

Figure 11. Top panel: Mean response latency in observed (Experiment 1) and simulated data, as the function of the proportion of congruent trials (25%, 50%, and 75%). Error bars represent 95% confidence intervals. Bottom panel: Example dynamics of conflict level (*C*) as well as control strength (*G*) in a typical simulated agent, as the function of the proportion of congruent trials (25% and 75%). In primarily incongruent sequences (25% congruent), both conflict level and control strength are permanently increased in comparison to primarily (75%) congruent sequences. A small decrease in values of both variables throughout the sequence is due to utility learning.

proportion-congruent effect, which cannot be explained solely by the conflict evaluation mechanisms (also see the Theoretical Contribution section).

Simulation 7 (item-specific proportion-congruent effect). One problem with explaining the proportion-congruent effect in terms of generally increased conflict in trials following incongruent trials stems from the observation of the item-specific proportion congruent effect (ISPC; Jacoby et al., 2003; Jacoby, McElree, & Trainham, 1999), that is, a decreased congruency effect for stimuli that appear more often in incongruent trials than for stimuli that appear less often in such trials. The ISPC effect is an example of a broader class of context-specific congruency effects (e.g., smaller congruency effects for locations that more often yielded incongruent stimuli; Crump, Gong, & Milliken, 2006). Such effects can be explained only by assuming some mechanism that encodes how much conflict each stimulus (or each location) has elicited so far, and how much increases control for the high-conflict stimuli. Such a mechanism is absent in the most influential Botvinick et al. (2001) model of conflict, and thus, that model is not able to replicate the ISPC effect.

In contrast, in the model of Verguts and Notebaert (2008), in which the RL algorithm was used in order to adapt control to locally defined conflict, the ISPC effect has been replicated successfully. The amount of conflict that was connected with a particular stimulus was encoded by means of the consequent strengthening/weakening of the task node associations in the case of stimuli yielding higher/lower conflicts, resulting in a faster/slower activation of responses to those stimuli (for another explanation of the ISPC effect in terms of the contingency between distractors and responses, see Schmidt, 2013).

Can the integrated utility-based model also generate the ISPC effect using its RL mechanism for rule utility? Unlike in the Verguts and Notebaert model, here the RL mechanism does not rely on the evaluated conflict in a direct way. Instead, positive feedback is simply assigned to the name-color rules if they yielded correct responses. Overall, such rules always yield correct responses. However, as in congruent trials, the read-word rule is often used to make the correct response (see Figure 4), there are fewer occasions to receive positive feedback for the name-color rule associated with a color that more frequently occurred in congruent trials, in comparison with the name-color rule associated with a color more frequently occurring in incongruent trials (given that an overall frequency of both colors in a sequence is comparable). In consequence, after several trials, the former rule surpasses the latter one in base utility, and thus it fires more quickly, yielding a lesser interference effect. The present solution of encoding local conflicts seems to be both simpler and more general than the Verguts and Notebaert solution because it requires sheer, not conflict-modulated, RL mechanisms (see also the Theoretical Contribution section).

The model successfully generated the ISPC effect observed in Experiment 2A from Jacoby et al. (2003), in which each of four stimuli appeared 48 times, but two stimuli appeared in 36 (75%) incongruent and 12 (25%) congruent trials, whereas the other two stimuli appeared vice versa. The simulated ISPC effect, compared to data from Jacoby et al. (2003), is presented in top panel of Figure 12. Bottom panel of Figure 12 presents the changes in the base utility of rules that responded to the 25% and 75% congruent stimuli.

Simulation 8 (preceding trial's congruency [Gratton] effect). A decrease in the congruency effect in trials following incongruent trials, in comparison with trials following congruent trials (Gratton et al., 1992; Kerns et al., 2004), can be explained by the neural networks (e.g., Botvinick et al., 2001) in a method similar to the

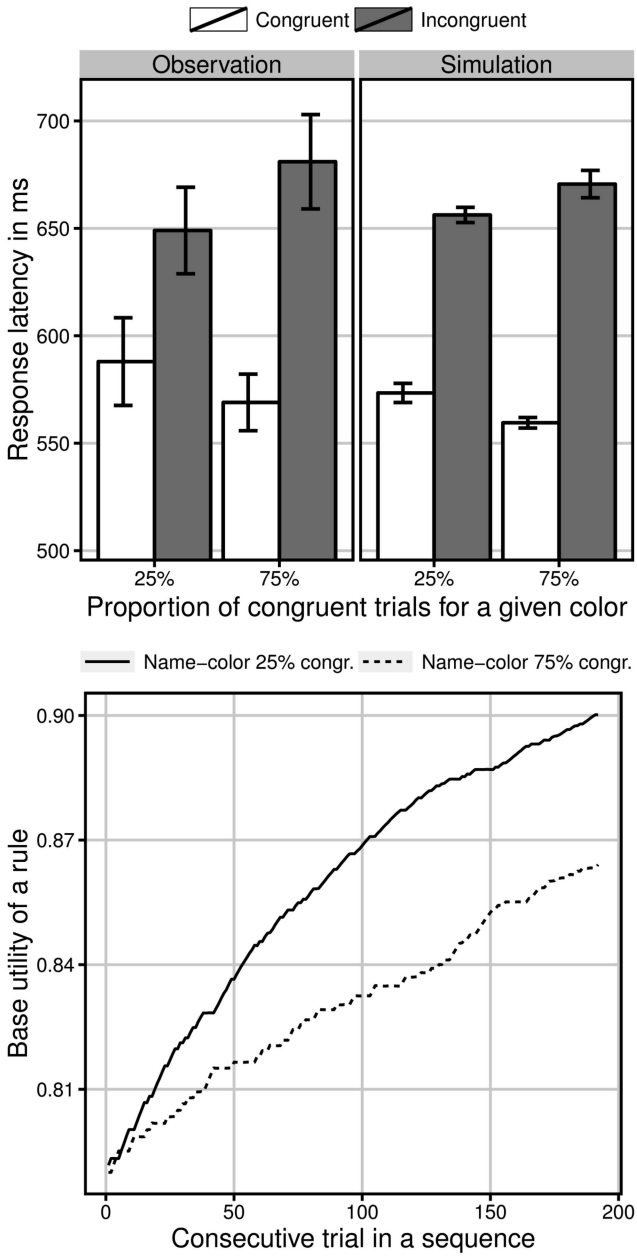

Figure 12. Top panel: Mean simulated response latency in the congruent and incongruent trials, as the function of whether a particular color occurred more (75%) or less often (25%) as congruent. Observed data from Jacoby et al. (2003; Experiment 2A). Error bars represent 95% confidence intervals. Bottom panel: Changes in utility of respective production rules during the task sequence.

proportion-congruent effect. Simply, an incongruent trial leads to an increase in control (due to detected conflict), and part of this control is passed to a subsequent trial (due to the control adaptation inertia). Our model relies on exactly the same mechanism in replicating the original Gratton data (see top panels of Figure 13). However, unlike the Botvinick et al. (2001) model that needed two different values of the learning rate parameter to replicate both the Gratton (high value) and proportion-congruent effects (low value; for discussion, see Jiang et al., 2014), our model accounted for both of these effects (as well as the ISPC effect), with no parameter modification (except for the time scale).

Analysis of the model dynamics in bottom panels of Figure 13 indicated that the pattern of latency data was matched by the patterns of conflict level and control strength. An increased amount of control passed from an incongruent, compared with a congruent, trial to a subsequent incongruent trial, and helped to choose the name-color rule in the latter trial. At the same time, the increased control prevented the read-word rule choice (which otherwise would speed up responding) in a subsequent congruent trial, making the latency in such a trial increase in comparison with a congruent trial preceded by a congruent trial. Thus, the congruency effect was visibly smaller in trials following incongruent, rather than congruent, trials.

Simulation 9 (N2/N450 and ERN waves observed in EEG research). It is commonly observed that Stroop-like tasks, especially the flanker task, yield two characteristic event-related potentials (ERP; EEG wave components appearing in a certain time window after stimulus or response onset). A negative peak in registered voltage, called N2, is observed about 250 ms–350 ms after an incongruent stimulus was presented, whereas no such wave (or a much lesser one) shows up after congruent trials (Kopp, Rist, & Mattler, 1996). N2 was observed primarily in the central/frontal electrodes, and is commonly associated with the anterior cingulate cortex (e.g., van Veen & Carter, 2002). Another negative peak, called error related negativity (ERN; Hohnsbein, Falkenstein, & Hoorman, 1989) is observed just after (or even just before) an incorrect response is made. It can also appear when the response

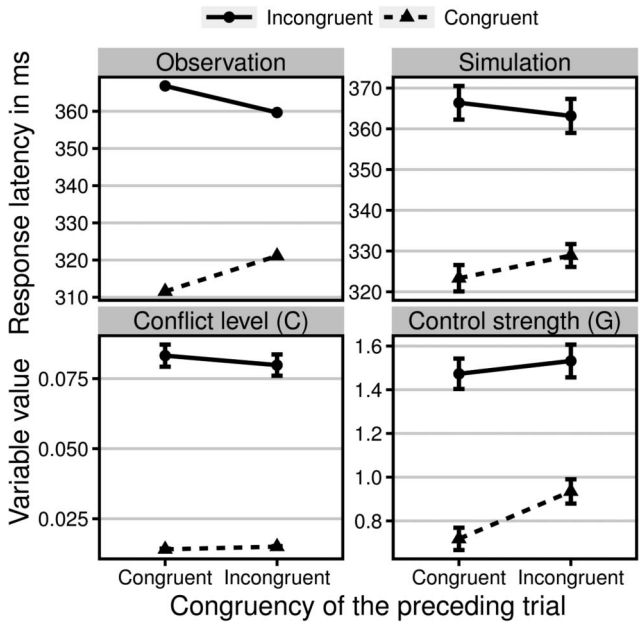

Figure 13. Top panels: Mean simulated response latency in the congruent and incongruent trials following either a congruent or incongruent trial, compared to data observed by Gratton et al. (1992; Experiment 1). Bottom panels: Mean values of the conflict level (*C*) and control strength (*G*) variables in the model, as the function of a preceding trial (congruent vs. incongruent). Error bars represent 95% confidence intervals.

is correct but different than predicted, slower than expected, or when negative feedback is presented (for a review, see Holroyd & Coles, 2002).

Conflict monitoring models predict that both N2 and ERN found in the flanker task reflect conflict detection. Whereas in the case of N2, this conflict has been effectively detected and resolved before a response is made, ERN indicates that the conflict has been detected too late, and has not been resolved before a response is made (see Yeung, Botvinick, & Cohen, 2004; Yeung & Cohen, 2006). Due to this fact, such a response will likely be incorrect. While the N2 explanation in terms of conflict is widely accepted (Yeung & Nieuwenhuis, 2009, but see Alexander & Brown, 2011), several alternative explanations of ERN were proposed, for example predicting that ERN results from a mismatch between a to-be-generated response and a response that is either intended (Scheffers & Coles, 2000) or predicted (Alexander & Brown, 2011; Holroyd & Coles, 2002; for an integrative account, see also Cockburn & Frank, 2011).

Although relatively scarce, the research on ERPs in the classical color-word task suggests that other EEG components may also be sensitive to conflict. One is N450, a negative wave that peaks about 450 ms after stimulus onset (for review, see Larson, Clayson, & Clawson, 2014), and due to that fact is unobserved in the flanker task, which yields responses shorter than 450 ms (e.g., Gratton et al., 1992; Yeung & Cohen, 2006). Similar to N2, N450 is larger for incongruent than for congruent stimuli, and often both show the similar Gratton effects (Larson, Kaufman, & Perlstein, 2009; West et al., 2012). In Simulation 9 we aimed to show an ability of our model to mimic conflict-related ERPs in correct trials (with the values of conflict *C* reflecting respective EEG amplitudes), without any univocal interpretation of such a wave as either N2 or N450 (see also Simulation 18). We also modeled the ERN wave.

In principle, the integrated utility-based model was designed to explain behavioral (accuracy and latency) effects found within the Stroop paradigm, and was not intended to describe their underlying neuronal level. However, due to rule "wait," the model usually chooses responses in more than one cycle (with the conflict level, control strength, and momentary utilities varying between cycles), so the model displays within-trial internal dynamics of conflict (i.e., the changes in variable *C*) that could be associated with ERP. In the present simulation, we tested whether such changes in *C* matched the commonly observed patterns of N2/N450 and ERN that have successfully been simulated by the conflict-monitoring networks (e.g., Yeung et al., 2004; Yeung & Cohen, 2006), as well as by the RL models (e.g., Alexander & Brown, 2011; Holroyd & Coles, 2002; Holroyd et al., 2005). If we are able to replicate the N2/N450 and ERN waves using our model, then its plausibility could be supported. This would also show that phenomena pertaining to neural dynamics can be replicated even if the model is formalized on the abstract level of rules and their utilities, and not directly on the neural level.

In order to generate the EEG waves, we used trials in which at least two cycles were carried out by the model. We recorded the initial level of variable *C* during stimulus presentation, intermediate level(s) captured between subsequent cycles, and the final level at response generation. Moreover, after a response was made, the model was allowed to run for several cycles (but no longer emitting any response), in order to observe changes in *C* after responding (i.e., ERN). Following Yeung et al. (2004), we assumed that ERN results from the continued evaluation of conflict even if the response has already been made. We found an analogue of the N2/N450 wave in the correct incongruent, but not in the correct congruent, trials, and not in the incorrect trials. There was also an analogue of the ERN in the incorrect trials but not in the correct ones. Figure 14 shows the successfully replicated ERP patterns (the changes in value *C*), matched to data from the flanker task presented in Yeung et al. (2004).

Simulation 10 (larger N2/N450 amplitudes for slower responses). Another crucial effect regarding ERP waves consists of a larger N2 amplitude in the flanker trials in which relatively slow latency accompanies higher accuracy, in comparison with faster trials that usually yield lower accuracy (Yeung & Nieuwenhuis, 2009). Yeung and Nieuwenhuis demonstrated that a variant of the Botvinick et al. (2001) model easily accounted for the above effect in the flanker task. We also attempted to replicate that effect in the case of the color-word task. We simply took the data from Simulation 9, and split all of the trials into two halves according to the median RT (727 ms). As expected, the faster half of the trials yielded lower accuracy (M = .94) than did the slower half (M = .96), $t(339) = 6.75$, $p < .001$. Most importantly, the slower trials yielded a substantially larger conflict, as reflected by the peak value of *C*, compared with the faster trials (see top panel of Figure 15). Moreover, West and Alain (2000) showed than in the PIS condition, the conflict elicited in fast trials was so low (as indicated by low N450) that the respective congruency effect in latency became insignificant, while in slow trials it was substantial (as also was N450). In order to check whether our model predicted such an effect, we took the fast and slow trials from the PIS condition of

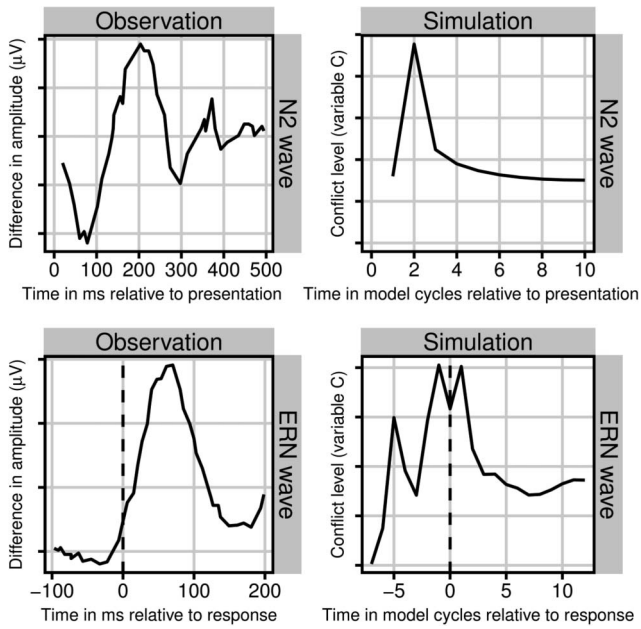

Figure 14. Top panel: Changes in the simulated conflict level (variable *C*) in the incongruent correct trials (the simulated N2 wave) over consecutive model cycles relative to presentation. Bottom panel: Changes in the simulated conflict level (variable *C*) in the incongruent incorrect trials (the simulated ERN wave) over model cycles relative to response. Compared to data from Yeung et al. (2004).

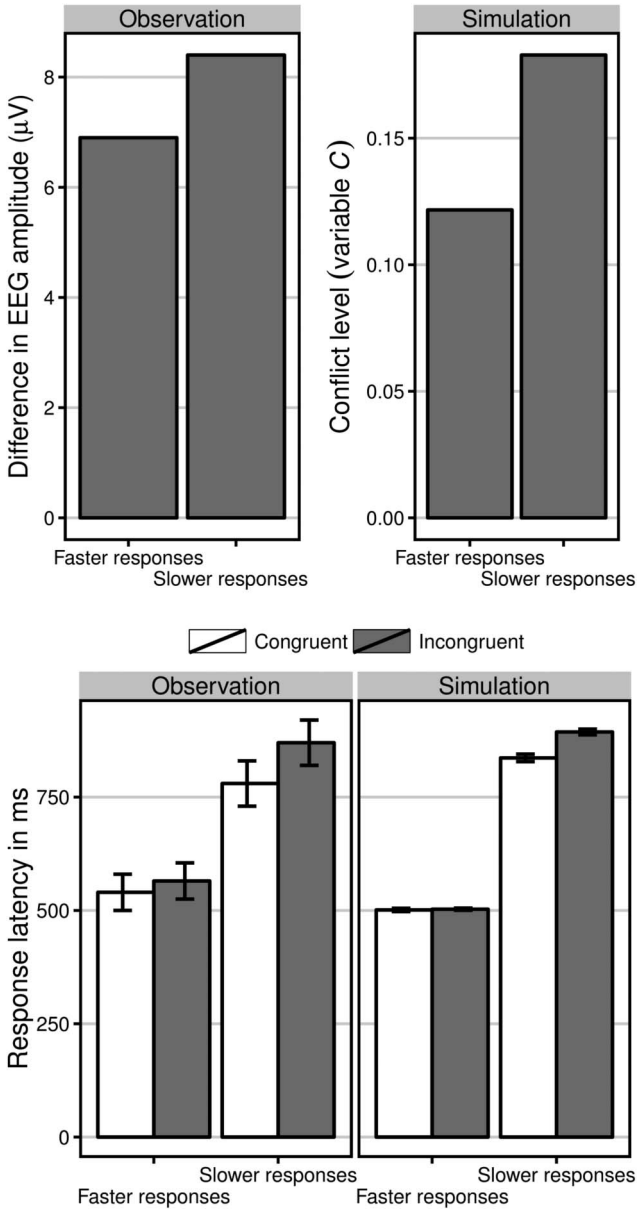

Figure 15. Top panel: The maximum difference in EEG amplitude between the incongruent and congruent trials below (faster responses) and above median RT (slower responses; Yeung & Nieuwenhuis, 2009), compared to the maximum conflict level (variable *C*) in the respective simulated trials. Bottom panel: The observed (West & Alain, 2000) and simulated mean response latency in the congruent and incongruent trials, for faster versus slower responses, in the primary incongruent sequence. Error bars represent 95% confidence intervals.

Simulation 6 and we observed (see bottom panel of Figure 15) the corresponding magnitude of the congruency effect, as did West and Alain. In particular, no reliable effect was found in the fast trials.

Stimulus-Related Effects

In this section, we replicated four effects related to the conditions of stimulus presentation: the temporal (stimulus onset asynchrony) and spatial disintegration effects, the smaller interference effect for the word-word and color-color tasks, compared with one in the color-word task, and the semantic gradient effect.

Simulation 11 (temporal disintegration [stimulus onset asynchrony] effect). Glaser and Glaser (1982; Experiment 1) demonstrated that the congruency effect was maximal when the target and distractor co-occurred, but almost disappeared when they were separated by an interval of more than 100 ms (see also Dyer & Severance, 1973). As noted, this effect could not be replicated by the neural networks (e.g., Cohen et al., 1990; Cohen & Huston, 1994; Phaf et al., 1990; Stafford & Gurney, 2007), as all of those models predicted substantial increases in the congruency effect with increasing SOA between words and colors (instead of its observed decrease). In contrast, Roelofs (2003) and Lovett (2001) successfully replicated this effect (see also Altmann & Davidson, 2001). In the present model, the effect was a direct consequence of the fact that the visual buffer contents determined which rules contributed to conflict. The postponement of a stimulus presentation resulted in, for some time, only one chunk (i.e., either word or color) occupying the visual buffer. In consequence, only rules that matched that one chunk could be included in the response set, as well as could be used in conflict evaluation and rule selection. This fact strongly reduced interference. Unlike in Roelofs (2003) and Lovett (2001), no parameter had to be modified in order to precisely fit the observed data (except for time-scalling parameters), because the effect simply resulted from the adopted architectural assumptions pertaining to the visual buffer. Figure 16 shows the comparable observed and simulated SOA effect. There was also a match regarding how SOA affected facilitation. In Glaser and Glaser (1982), this effect was constant for all values of the negative SOA (distractor comes first) but disappeared under positive SOA (target comes first), and this pattern of data was also found

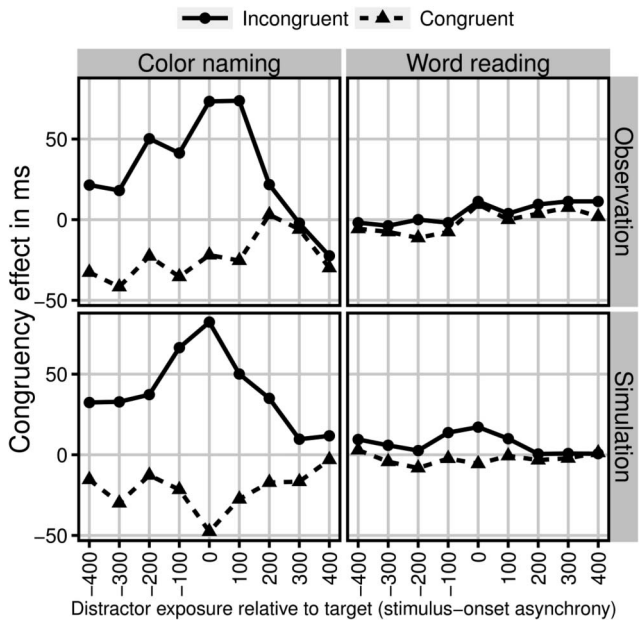

Figure 16. Interference and facilitation effects for different values of stimulus onset asynchrony, for the color naming and word reading tasks. Observed data from Glaser and Glaser (1982, Experiment 1).

in our data. Notably, similar to the Glaser and Glaser data, no substantial SOA effect was found in the case of word reading. However, a small but significant interference of 36 ms was simulated for SOA = 0 ms (note that some amount of this reverse interference can be observed for all of the SOA levels except for SOA greater than 200 ms), whereas Glaser and Glaser observed only a marginally significant effect of 8 ms. Further discussion of this reverse interference effect is presented in Simulation 16.

Simulation 12 (spatial disintegration effect). Another important stimulus-related effect consists of a decreased but significant interference when the color and the word are separated in space (i.e., a stimulus is disintegrated; Dyer, 1973; Kahneman & Henik, 1981), compared with when the integrated, colored word is presented. In the present model, the effect directly resulted from architectural assumptions regarding the visual buffer. The spatial separation was accounted for by placing a color and a word chunks in two distinct locations (hand-coded) of the virtual screen. As the model could focus on only one (randomly chosen) location at a time, different sets of conflicting rules (i.e., "name color" and "other") were selected when the model "looked" at a color compared with when it focused on a word (i.e., "read word" and "other"). When the model looked at a color, little conflict was elicited and interference was low. When it looked at a word, the task could not be accomplished (no color to name), and the model required attention switch to the location of a color in order to eliminate the read-word rule from the response set and to include the name-color rule instead. The model predicted that interference decreased by around 50% because on half of the trials (i.e., when the model "looked" at a word), a time-consuming switch to a color was needed; on the other half of the trials (when the model directly "looked" at a color), colors could be processed immediately and without conflict, yielding no additional latency. The mean interference effect was thus the mean of these two cases (see Figure 17). As noted by a reviewer, this fact implies a prediction to be tested in the future that under spatial separation an approximate binomial distribution can be expected. However, the fit to data was not ideal: the simulated effect was smaller than the one observed by Dyer (1973).

Simulation 13 (smaller word–word and color–color interference). Congruency effects are commonly observed when two incongruent stimuli of the same type activate the same process, like in the so-called word-word and color-color tasks (for a review, see MacLeod, 1991). For example, in a classic study, Dallas and Merikle (1976) presented one word below/above another word or a nonword. A cue presented 250 ms before/after the presentation of stimuli indicated the target word to be read. The authors found a larger latency of response for the word than for the nonword distractors if the cue followed the stimuli. The size of the interference was relatively small (46 ms), compared with the usual interference in the color-word task (>100 ms). Weaver++ successfully replicated this effect (see Figure 17 in Roelofs, 2003). Also, the replication by Lovett (2005; see her Figure 2) of the moderate interference for both equally learned tasks from the MacLeod and Dunbar (1988) experiment suggests that her model would generate this effect. Only the Cohen et al. (1990) model was unable to yield interference between two processes of equal strength. The present model easily explains why reduced but significant interference should be expected in the word-word task (the same logic applies to the color-color task). Simply, due to

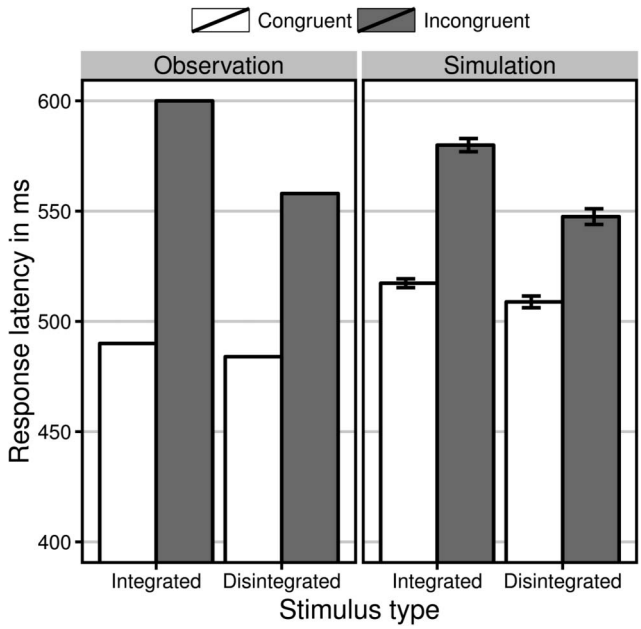

Figure 17. Observed and simulated mean response latency for stimuli that are either spatially integrated (observed data from Dyer, 1971, SOA = 0 ms) or disintegrated (observed data from Dyer, 1973). Error bars represent 95% confidence intervals.

formula (2), with two mutually incongruent read-word rules equal in utility ($U = 1.0$), the one responding to a distractor occupies the numerator of the formula, while both are placed in the denominator. So, the conflict level is around 50% (1.0/[1.0 + 1.0]). When the most goal-relevant rule has a relatively lower utility ($U = 0.7$), as occurs in the color-word task, then the conflict will increase to about 60% (1.0/[0.7 + 1.0]), and usually its resolution will last longer, as a larger increase in control strength will be needed to overcome the goal-irrelevant rule and to select the proper rule. Figure 18 summarizes the decrease in interference when the utility of the most goal-relevant rule gradually approaches the goal-irrelevant rule's utility equaling unity, compared to our own data on the color-word ($U = 0.79$) and figure-word interference ($U = 0.90$) from Experiment 1, as well as to Dallas and Merikle's original data on word-word interference ($U = 1.0$).

Moreover, Glaser and Glaser (1982; Experiments 3 and 4) found that when the color-color task informed people (i.e., under spatial certainty) which of two mutually incongruent color patches separated in space was relevant, the congruency effect became insignificant, in contrast to the spatial uncertainty condition (see Simulation 12). Only the Roelofs (2003) model successfully simulated this effect, although Roelofs needed to manipulate one crucial parameter reflecting the strength of control over distraction (parameter *du*). The integrated utility-based model replicated the spatial certainty effect in the color-color task analogously as it accounted for the spatial disintegration effect, with no parameter modification, just by placing two color chunks, instead of a word and a color, in two distinct locations of the visual buffer (both evoking a name-color rule), and always setting the focus on the relevant location. When the model "knew" which color to focus on and did not encode the irrelevant color in the visual buffer, the

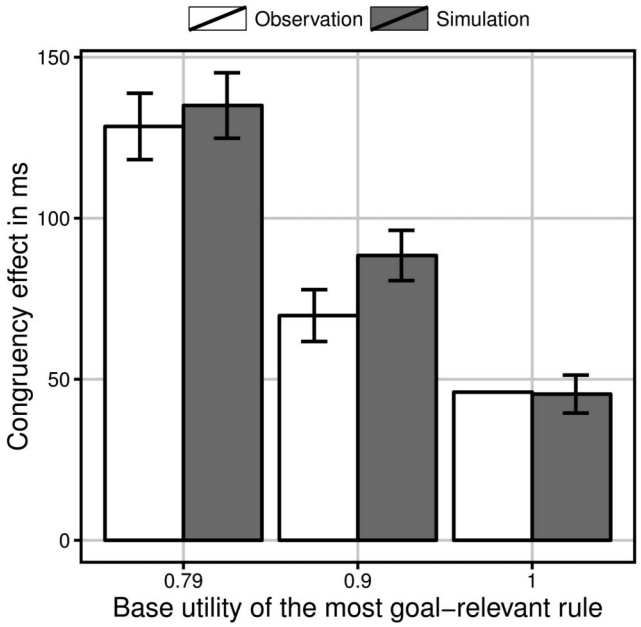

Figure 18. Simulated interference effect for consecutive values of the most goal-relevant rule base utility (when the alternative rule has base utility of $U = 1.0$), compared to effects observed in the color-word (left white bar; data from Experiment 1), the figure-word (middle white bar; data from Experiment 1) and the word-word tasks (right white bar; data from Dallas & Merikle, 1976, Experiment 1). Error bars represent 95% confidence intervals.

interference was minimal in comparison with when the location initially attended to was random, and when there was 50% chance of encoding a distractor. Figure 19 presents simulation results matched to Glaser and Glaser's data.

Simulation 14 (semantic Gradient effect). Interference in the color-word task is highly reduced, but is still significant, when the distractor word does not denote a color, but instead names another entity semantically related to that color (e.g., an object that usually appears in a particular color; Dalrymple-Alford, 1972; Klein, 1964; for reviews see MacLeod, 1991; Roelofs, 2003). For example, naming the red color of the word BANANA yields a larger latency than is yielded by nonsense syllables (although it is smaller than latency yielded by naming the red color of the word YELLOW). The more a word is semantically connected with a particular color, the larger the interference effect. The semantic gradient effect was also demonstrated in other Stroop-like tasks, like the picture-word task (La Heij & van den Hof, 1995). Roelofs (2003) explained the effect in terms of the automatic spreading of activation in semantic memory, from a distractor to features associated with that distractor, including its color. Such a memory activation of the incongruent color makes the correct color representation retrieval more difficult due to competition, and delays that retrieval. An analogous mechanism was also implemented in the Lovett (2005) model (see also Altmann & Davidson, 2001). No simulation of the gradient effect has ever been attempted in the case of neural networks.

When we assumed that a color-associated word activated rule "name object's color" instead of the read-word rule (as words no longer named colors), the semantic gradient effect could easily be replicated in the present model, despite not accounting for linguistic processes that occur in the Stroop. As naming usual colors of particular objects is quite a rare action, obviously less practiced than is naming sheer colors, in the simulation we assigned to rule "name object's color," which was activated by object names included in the task, the base utility of $U = 0.25$, which is substantially lower than $U = 1.0$ for the read-word rule. The observed and predicted latencies for naming the colors of color words versus the colors of color-associated objects, compared to data from Dalrymple-Alford (1972), are presented in Figure 20. Note that the fit to data was not perfect: the simulated effects were smaller than (exceptionally large) observed effects.

Response-Related Effects

In this section, we describe two classic effects related to manipulations made to the response conditions: presenting distractors associated with no response (the response-set effect) and changing the task to word reading (reverse interference). Moreover, two novel predictions that result from adopting the Festinger conflict formula (but not expected under the Hopfield formula) are analyzed. One prediction holds that the total number of possible responses in the task does not affect the congruency effect (no response-set size effect), which Chuderski, Smoleń, and Taraday (2014) recently confirmed. The other predicts that the number of possible responses for a distractor has a great impact on conflict, which Chuderski, Senderecka, Kałamała, Kroczek and Ociepka (2015) recently corroborated.

Simulation 15 (response-set effect). A decreased interference is observed when a distractor word does not name any color presented in the task (Proctor, 1978). In our model, this effect

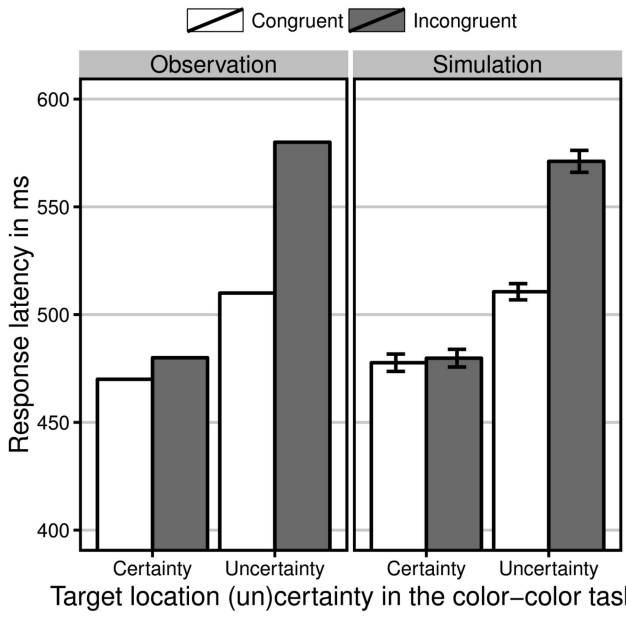

Figure 19. Simulated mean response latency for the incongruent versus congruent trials of the color-color task, matched to data observed by Glaser and Glaser (1982) under spatial uncertainty (Experiment 3) and spatial certainty (Experiment 4). Error bars represent 95% confidence intervals.

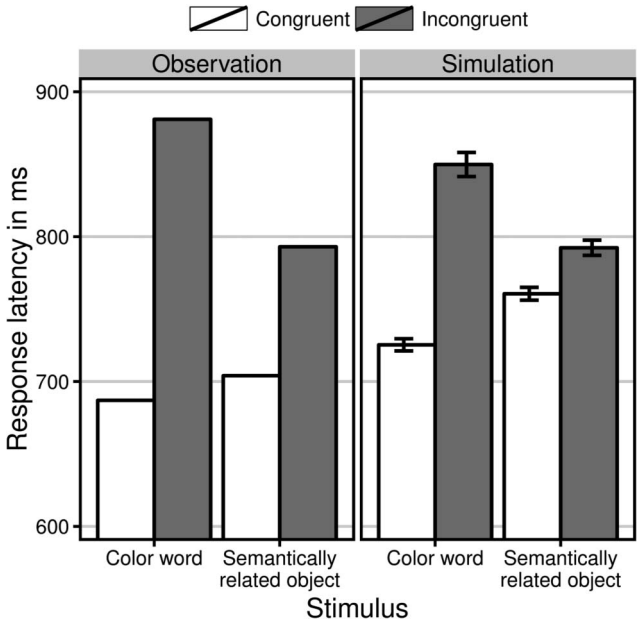

Figure 20. Observed (Dalrymple-Alford, 1972) and simulated mean response latency for distractors being either a name of color (e.g., YELLOW) or an object semantically related to that color (e.g., BANANA). Error bars represent 95% confidence intervals.

results from the fact that the rules that associate particular colors (either their names or corresponding words) with the respective manual responses (i.e., button presses) are available only for those stimuli that have some button assigned, and could be compiled during training. Hence, there can be no specific rule for a word that names a color that has not been defined in the task instruction. In consequence, when a word from outside the set of possible responses is presented, only two rules are activated: "name color" and "other." Because the conflict level is lower (no read-word rule has to be overcome), compared with when a word elicits a response, the response latency is substantially lower than in the case of distractors from inside the set of possible responses (see Figure 21).

Simulation 16 (small but nonzero reverse interference effect). Although early research indicated no interference from colors on word reading, that is, no reverse interference effect (e.g., Glaser & Glaser, 1982; but see Dyer & Severance, 1972; Stroop, 1935), and consequently, the existing Stroop models predicted a lack of such an effect (e.g., Cohen et al., 1990, Figure 5; Herd et al., 2006, Figure 4; Lovett, 2001, Figure 1; Roelofs, 2003, Figure 13; van Maanen et al., 2009, Figure 1; for one exception see Phaf et al., 1990), the existence of a small but significant reverse effect has been noted in more recent studies (for a review see Blais & Besner, 2006). In some studies, the processing of words was made more difficult (due to their rotation, distortion, small size etc.), and in some others, the response to words required the translation of a linguistic representation to another response code (e.g., sorting words to bins labeled with color patches), so the real existence of the reverse effect could be questioned. Blais and Besner (2007), however, in a methodologically improved experiment, demonstrated a reliable interference effect of 16 ms from colors in the word categorization task. We tested whether our model also generated nonzero interference when its goal was to "read words." Indeed, in line with Blais and Besner (2007), the model generated small but significant interference of 36 ms in the word reading task (see Figure 16).

Simulation 17 (no response-set size effect). It has long been debated whether increasing the size of the set of colors (and corresponding responses) in the Stroop task increases or decreases the congruency effect. For example, MacLeod (1991) cited several studies published between the 1960s and the 1980s that showed no set-size effects, while three studies yielded a decrease in interference and another three yielded an increase. Two relatively newer studies (La Heij & van den Hof, 1995; Kanne, Balota, Spieler, & Faust, 1998) also provided incoherent evidence. La Heij and van den Hof (1995), using the picture-word task, compared conditions including four versus 16 pictures and found an increase in interference as a function of the response-set size. However, increasing the set so much might have also additionally loaded working memory, possibly causing more interference. Kanne, Balota, Spieler, and Faust (1998), using the color-word task, found an increase in interference only from two to three colors, but not from three to four. The two-stimuli condition is special, because each color is perfectly correlated with a distractor word (see Melara & Algom, 2003), so it should not be taken into account. Thus, Kanne et al.'s (1998) results seem to suggest no response-set size effect.

The linguistically oriented Stroop models (e.g., Altmann & Davidson, 2001; Roelofs, 2003; van Maanen, van Rijn, & Borst, 2009) predicted substantial such an effect (see Roelofs, 2001) as a direct consequence of adopted mechanisms of chunk retrieval. These models assumed that one and the same process generates output for color naming and word reading, and the interference

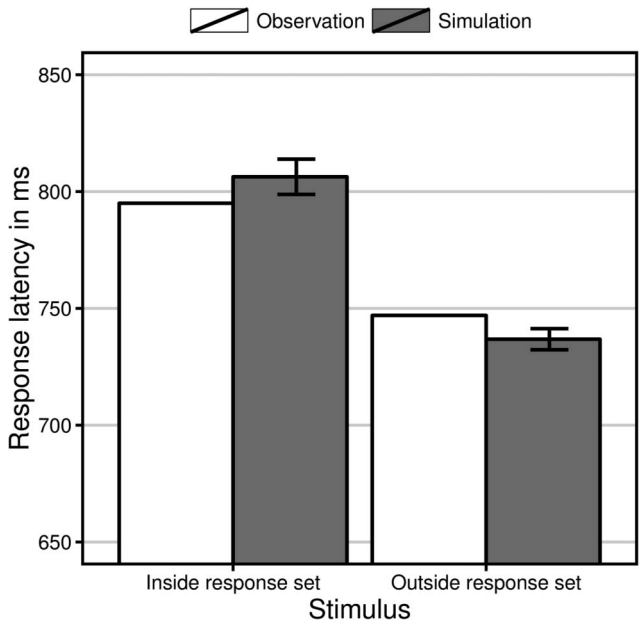

Figure 21. Simulated mean response latency for color words that either do (inside response set) or do not (outside response set) name a color allowed in the task. Observed data from Proctor (1978, Experiment 3). Error bars represent 95% confidence intervals.

effect resulted from the need to retrieve from memory the proper representation of a response (e.g., a lemma triggering an utterance; Roelofs, 2003). Chunks related to word reading were more available, so their retrieval was more probable, but the goal representation boosted the activation of the color-related chunks, and allowed for their retrieval, though yielding an additional processing cost. The key attribute of these models was that retrieval latency of the proper color-related chunk was the inverse function of the ratio of its activation to the summary activation of all potential distractors (respective word-related chunks). Thus, in incongruent trials, the larger the response-set size, the more distractors compete for retrieval (i.e., the more increases the denominator of the ratio, thus the ratio decreases), and the longer it takes to retrieve the color-related chunk. Because neither neutral nor congruent trials activated distractors, thus the linguistically oriented models predicted the increase in the congruency effect as the response-set size increased (Roelofs, 2001; a similar effect to be found in Lovett, 2005). In addition, the Cohen et al. (1990) model predicted a small but systematic increase with increasing response-set size (see Cohen, Usher, & McClelland, 1998).

In contrast, the present model predicts the absence of a response-set size effect as a consequence of the fact that only rules applicable for a given target and distractor are considered in the response set. Thus, the conflict level is constant (ceteris paribus) no matter how many name-color and read-word rules are present in the model. Recently, Chuderski et al. (2014) tested this prediction with regard to the manual variant of the color-word and figure-word tasks, using a design that reduced WM load in the task, and controlling for factors that co-occurred with the increase in response-set size, like the number of trials per color or per target-distractor pair. As predicted, they found no statistically significant difference in the congruency effect between the conditions that required four, six, and eight targets/responses. Figure 22 compares observed and predicted data, the latter also yielding an insignificant effect, $F(2, 1017) = 0.035$.

Simulation 18 (the number of applicable responses effect). Here, we tested the crucial prediction of the adopted Festinger formula of conflict evaluation, which is related to a situation when more than one response can be validly undertaken in reaction to a certain stimulus (either target or distractor). As already discussed in the text and as demonstrated in Figure 3, our calculations suggested that the alternative Hopfield formula of conflict evaluation could not handle the case when the number of applicable responses for a distractor increased. In contrast, that case was nicely handled by the Festinger formula, which predicted a sharp increase in conflict.

In our recent EEG experiment (Chuderski et al., 2015), we directly tested this prediction, by measuring the brain correlates of conflict (for their discussion see Simulation 9) in a Stroop task that varied the number of allowed responses for particular colors. Specifically, we used three colors and three distractor words naming these colors. Participants responded to colors with six response keys, using three fingers of each hand. Randomly, one color was associated with one possible key, the second—with two alternative keys (i.e., participants could use any of them to respond), and the third—with three alternative keys. For each color, there was one congruent word meaning that color, and thus colors (X-) and words (-X) primed the same response(s) (Conditions 1–1, 2–2, and 3–3).

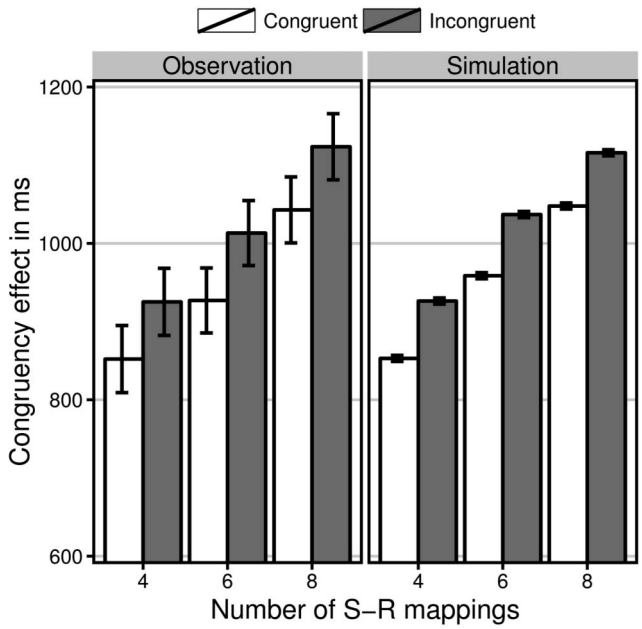

Figure 22. Observed (Chuderski et al., 2014, Experiment 1) and simulated mean response latency in the congruent and incongruent trials for four, six, and eight possible stimuli/responses in the color-word task. Error bars represent 95% confidence intervals.

Also, for each color there was one incongruent word that primed more (incorrect) responses (Conditions 1–3, 2–3, and 3–2, for 1, 2, and 3 correct responses, respectively) than did the other incongruent word (Conditions 1–2, 2–1, and 3–1, respectively). The Festinger formula predicts a higher level of conflict for the former conditions, in comparison with the latter conditions. All stimuli were fully randomized in a sequence of 192 congruent and 240 incongruent trials. Other details of this study were analogous to Experiment 1.

We recorded the N450 wave (in the 400 ms–540 ms window) in 33 people performing the task. Our results clearly confirmed the predictions. For each color, the N450 amplitude, recorded at fronto-central electrodes, was more pronounced (i.e., more negative) in the incongruent trials priming more possible incorrect responses than in the incongruent trials priming less such responses. An increased number of incorrect responses primed by a distractor word yielded, on average, a 0.52 μV drop in amplitude, $t(32) = 2.19$, $p = .018$, Cohen's $d = 0.17$, one-tailed test. The results for electrode CP2, where the effect was most visible, are shown in Figure 23. Finally, we simulated these data, using the same procedure as in Simulation 9. We observed the resulting dynamics of conflict variable *C*. The model predicted a larger overall conflict for the higher number of possible responses to distractors, compared with the lower number. In total, these results support the Festinger formula for conflict evaluation as more plausible than the Hopfield formula.

General Discussion

The integrated utility-based model successfully explained 18 experimental Stroop effects, being the largest data set explained to

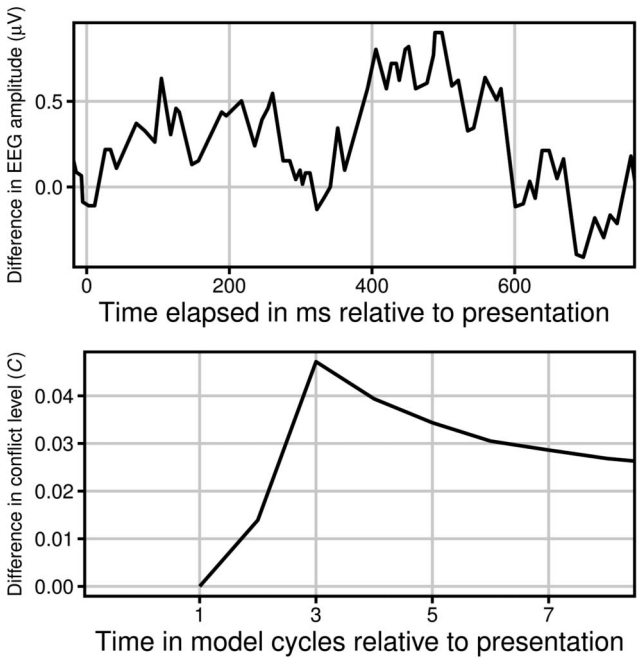

Figure 23. Observed (Chuderski et al., 2015) difference in EEG amplitude between the incongruent trials in which a relatively lower number of responses was primed by distractors and the incongruent trials in which a relatively higher such number was primed (i.e., the difference reflects the amount of additional negative deflection of N450 in the latter trials), compared to an analogous difference in the simulated conflict level (variable *C*), shown over consecutive model cycles relative to presentation.

date by one Stroop model. These included basic congruency effects, performance dynamics and adaptation, experimental manipulations to stimulation, and manipulations to responding.

Comparison With Alternative Models of Stroop

The integrated utility-based model shares certain characteristics with all three leading categories of Stroop models: the conflict-monitoring neural networks (Botvinick et al., 2001), the language production system Weaver++ (Roelofs, 2003), and the utility-learning model implemented in ACT-R (Lovett, 2005), but at the same time, it introduces several completely novel mechanisms. With the neural networks, the present model shares the idea of conflict evaluation, absent in both Weaver++ and the utility-learning model. However, it defines the amount of conflict as the *ratio* of competing responses to all applicable responses (Festinger's definition), in contrast to Botvinick et al.'s (2001) definition of conflict as the *product* of competing responses (Hopfield energy). With Weaver++ and the utility-learning model, the present model shares the production rule architecture, including visual attention that helps to account for the stimulus-related effects. Unlike these models, the integrated utility-based model abstracts from language production and semantic memory. Paraphrasing La Heij and van der Hof (1995), it predicts that the lion's share of the Stroop effect lays in response competition, whereas conflicts during language production and/or memory retrievals add minimally to this effect. With the utility-learning model, the present model shares the mechanism of utility reinforcement learning, and a substantial focus on the role of rule utility in response selection. However, only the present model directly integrates rule utility calculation with conflict evaluation and control regulation into one cognitive mechanism resolving the Stroop interference.

Table 2 presents a summary of the fits of the integrated utility-based model, compared to the effects fitted by the competing models. The integrated utility-based model successfully accounted for all of 18 selected effects, doing substantially better compared with the competitors, as each of them was able to explain only eight effects. Moreover, unlike the alternative models, the present model explained four crucial effects (Simulations 7, 8, 11, and 12) without assuming any additional parameter/mechanism. Below, we summarize the mechanisms that led to replication of particular groups of effects, as well as discuss reasons for the inability of competitive models to account for certain effects.

Each alternative model nicely accounted for two basic effects pertaining to interference and facilitation: the former exceeding the latter (Simulation 1), and the reduced interference in the picture-word task (Simulation 4). However, none of the competitors accounted for the distribution of response latency in congruent and incongruent trials (Simulation 2), or for the distributions of individual congruency effects and accuracy (Simulation 3). As shown by Mewhort et al. (1992), the Cohen et al. (1990) model generated an incorrect latency distribution, since in incongruent trials it predicted increases in either the central tendency or in the tail, but not both at the same time, whereas this latter pattern is commonly observed (Experiment 1; Mewhort et al., 1992; Roelofs, 2012). Neither Weaver++ nor the utility-learning model attempted to generate the latency distributions in Stroop, but in principle, they are able to give precise chronometric predictions, and in the future, it might be interesting to test what distributions they generate. In summary, the correct replication of RT distribution strongly speaks in favor of the integrated utility-based model.

The practice effect (Simulation 5) is a large problem for the neural networks. In particular, the Cohen et al. (1990) model incorrectly replicated the MacLeod and Dunbar (1988) data, as interference, present in these data when both tasks (the trained and the untrained) equaled in strength, completely disappeared in the simulation. This fact suggests that the implementation of practice in terms of increased strength of association between stimulus and response, assumed by the neural networks, is insufficient to explain the effects of practice. In contrast, both the utility-learning model and the present model assume that it is the actions' utility for the goals of an agent that matters in response selection, and underlies practice. The correct replication of the practice effect by these models suggests that explanation of action selection in terms of response-outcome, instead of stimulus-response, associations is more plausible.

Weaver++ takes a different solution in accounting for the MacLeod and Dunbar data. It assumes that extensive practice results in architectural changes in the trained process. Initially mediated by the conceptual (semantic) memory access, after practice, this process became compiled into the direct stimulus-response link. Although this solution allowed Roelofs (2003) to correctly replicate the practice effect, it seems problematic because in many cases learning becomes effective too quickly to yield any permanent changes in the brain structure. For example, the ISCP effect (Simulation 7) is correctly predicted by utility learning in the

Table 2
Summary of the Stroop Phenomena Replicated by the Integrated Utility-Based Model as Well as the Alternative Stroop Models

Effects	Sim. no.	Fig. no.	Effect name	Observed in	Neural networks	Weaver++	Utility-learning model	Integrated utility-based model
Basic congruency effects	1	4	Interference exceeds facilitation	Stroop, 1935; Dyer, 1971	+	+	+	+
	2	5, 6	Congruent/incongruent trials latency distribution	Mewhort et al., 1992; Experiment 1	—			+
	3	7, 8	Congruency effect distribution	Experiment 1				+
	4	9	Smaller congruency effect in the picture-word task	Hentschel, 1973; Experiment 1	+	+	+	+
	5	10	Practice reverses interference	MacLeod & Dunbar, 1988	—	+	+	+
Dynamics and adaptation	6	11	Proportion-congruent effect	Logan & Zbrodoff, 1979; Experiment 2	+		+	+
	7	12	Item-specific proportion-congruent effect	Jacoby et al., 2003	+			+
	8	13	Preceding trial's congruency (Gratton) effect	Gratton et al., 1992	+			+
	9	14	N2/N450 and ERP waves found in EEG research	Kopp et al., 1996; Hohnsbein et al., 1989	+			+
	10	15	Larger N2/N450/SP amplitudes for slower responses	Yeung & Nieuwenhuis, 2009	+			+
Stimulus-related	11	16	Temporal disintegration (SOA) effect	Glaser & Glaser, 1982	—	+	+	+
	12	17	Spatial disintegration effect	Dyer, 1973		+		+
	13	18, 19	Smaller word-word/color-color interference	Dallas & Merikle, 1976	—	+	+	+
	14	20	Semantic gradient effect	Klein, 1964; Dalrymple-Alford, 1972		+	+	+
Response-related	15	21	Response-set effect	Proctor, 1978	+	+	+	+
	16	16	Small but non-zero reverse interference	Stroop, 1935; Blais & Besner, 2007	—	—	—	+
	17	22	No response set-size effect	Chuderski et al., 2014	—	—	—	+
	18	23	The number of applicable responses effect	Chuderski et al., 2015				+
			Total number of effects replicated:		8	8	8	18

Note. + = effect replicated correctly; — = effect replicated incorrectly; blank = effect not attempted in any available publication. Neural networks = models described in Cohen et al. (1990, 1998); Cohen and Servan-Schreiber (1992); Cohen & Huston (1994); Botvinick et al. (2001); Jones et al. (2002); Yeung et al. (2004, 2006); Stafford and Gurney (2007); Verguts and Notebaert (2008), and Davelaar (2008). Weaver++ = Roelofs (2000, 2001, 2003), Altmann and Davidson (2001), and van Maanen et al. (2007, 2009). Utility-learning model = Lovett (2001, 2005). For description of all effects and models see the text. SOA = stimulus-onset asynchrony.

integrated utility-based model, while it is unlikely that this effect is accompanied by structural changes in brain. Thus, although brain research indeed suggests structural changes in processing paths due to long-term practice (for review see Green & Bavelier, 2008), in the context of the Stroop task, the mechanisms of utility learning seem to account for learning effects in a more plausible way.

The effects related to conflict/control dynamics and adaptation clearly lead to the rejection of both Weaver++ and the utility-learning model in their present form. Out of five effects analyzed (Simulations 6–10), only one effect—the proportion-congruent effect (Simulation 6)—was claimed to be replicated by Lovett (2005; though she did not report her data). Simply, both of those models would unlikely account for dynamics and adaptation because they lack any mechanisms for conflict evaluation. In the utility-learning model, any effect of either momentary or prolonged increases in conflict could only be reflected by the between-trial changes in rule utility (Lovett used this mechanism in order to account for the proportion-congruent effect). However, even if utility learning was helpful in producing some prolonged adaptation effects in that model, it seems to be too slow, and thus insufficient, to account for the immediate increases in control when conflict momentarily increases (e.g., the Gratton and ERP effects). In contrast, the conflict evaluation mechanisms included in the neural networks and the integrated utility-based model naturally led to all effects of dynamics and adaptation. In particular, in each model, fast changes in evaluated conflict could be aptly mapped onto corresponding dynamics of brain activity, as tapped by the respective ERPs. Overall, the five dynamics and adaptation effects provided strong support for the integrated utility-based model, but also for the neural networks, especially as the latter provided an even more detailed account of the conflict-related ERPs (Yeung et al., 2004; Yeung & Cohen, 2006) than did the present model.

The stimulus-related effects were disastrous for the neural networks, as they generated increasing, instead of decreasing, interference resulting from negative SOAs (see Simulation 11), in addition to predicting no word-word interference at all, instead of a moderate one (Simulation 13). It is also unlikely that this class of models will replicate the spatial disintegration effect (Simulation 12), as well as the semantic gradient effect (Simulation 14), both not yet attempted by them. All of these problems result from the fact that the neural models lack any dedicated mechanisms of attention (a network fragment for perception has the same form as the fragment for responding). In contrast, the rich attentional mechanisms, included in Weaver++ as well as in the utility-learning model (the latter inherited from the underlying ACT-R architecture), allowed these models to nicely account for the stimulus-related effects. A conceptually similar mechanism (the visual buffer) was implemented in the integrated utility-based model, helping this model to correctly replicate all of these effects.

The final group of effects—the ones related to conditions of responding—is critical for the evaluation of competing models. Only the response-set effect (Simulation 15) was accounted for by each competing model. In line with the established data (Blais & Besner, 2007), solely the integrated utility-based model correctly predicted a small but significant reverse interference effect (Simulation 16), whereas each alternative model predicted no such effect. Moreover, in contrast with reliable data from Chuderski et al. (2014), nicely fitted by the present model (Simulation 17), all alternative models incorrectly predicted the substantial increase in interference when the number of stimuli/responses (response-set size) increased. Most importantly, a novel and highly specific prediction of the present model was confirmed, which pertained to the situation in which more than one response could potentially be made toward a distractor (Simulation 18). A larger interference was shown when more, than less, responses were associated with a distractor (i.e., there was more total utility "supporting" incorrect responses). No alternative model has attempted this latter prediction. In Figure 3, we demonstrated that the neural networks would likely generate the opposite (incorrect) pattern of data. It also is unclear how this effect could be accounted for by Weaver++ and the utility-learning model. Thus, the effects from Simulations 16–18 could be explained only by the integrated utility-based model, but not with any other competing model.

It needs to be underscored that each alternative class of models predicted several additional effects that were not analyzed in the present paper. For example, Lovett (2005) correctly predicted data from the double-response Stroop, which the present model did not attempt. However, such a variant of the Stroop task is highly unnatural (it is a dual task possibly involving different cognitive processes than the standard Stroop), prone to strategic influences (as evidenced by Lovett herself), and rarely used (for exceptions see Klein, 1964; Shimada & Nakajima, 1991). Thus, it does not seem to be crucial for explaining Stroop. In another ACT-R based model, van Maanen et al. (2009) replicated the discrepant effects of using the psychological refractory period paradigm (another example of dual tasking) on the color-word versus the picture-word interference, not accounted for by our model, either. In a conceptually similar ACT-R model (Juvina & Taatgen, 2009), several priming effects were modeled. The present model, lacking declarative memory necessary to account for such effects, is unable to replicate them. However, in Experiment 2, we showed that in the manual variant of the Stroop task, there was no negative priming effect at all, so priming effects may be important only for Stroop variants that heavily load on the memory/linguistic components, but are not so crucial for the core mechanisms coping with Stroop interference.

The neural networks (e.g., Yeung et al., 2004; Yeung & Cohen, 2006), including those that modeled the RL mechanisms (Alexander & Brown, 2011; Holroyd et al., 2005), generated very specific effects related to EEG waves. As the present model was not designed at the neural level, and it primarily aims to account for behavioral effects, it may be not able to replicate all effects found on the neural level. Nevertheless, the model did generate the overall pattern of N2/N450 and ERN waves, a larger N2/N450 amplitude for slower/more accurate responses, as well as a larger N450 amplitude for trials in which relatively more responses were associated with distractors (see Simulations 9, 10, and 18, respectively).

Weaver++ (Roelofs, 2003) simulated several linguistic effects. We demonstrated that one such effect—the semantic gradient—could be explained just on the basis of differences in rule utility (Simulation 14), without appealing to linguistic processes. Analogously, a smaller between- than within-language congruency effect in bilinguals, modeled by Weaver++, could easily be modeled by accepting a rational assumption that reading words in a second language had a lower utility than did reading them in the

mother tongue (but a higher utility than color naming in the latter language). Although we do not deny that linguistic processes may affect interference in the cognitive system, we predict that the core mechanism of interference generation in the Stroop relies on more universal response selection mechanisms.

The increase in interference in aging, frontal lesions, schizophrenia, and dementia, which have been simulated by Weaver++ (Roelofs, 2003), the utility-learning model (Lovett, 2001), and neural networks (Cohen & Servan-Schreiber, 1992), by means of weakening the attention focus on targets, the activation spread from goal to memory chunks, or the influence of task-node on the network, respectively, can analogically be replicated in the present model by rationally assuming that in all of those cases the maximal strength of control (parameter *g*) is highly decreased (see Figure 8).

Finally, because we were primarily interested in the mechanisms of response selection in the Stroop, we did not address manipulations related to low-level perception of Stroop stimuli (e.g., words legibility and rotation), nor the dynamics of attentional focus coping with such stimuli, which, for example, were accounted for by the neural networks (Cohen et al., 1990), as well as the diffusion (White, Ratcliff, & Starns, 2011) and Bayesian models (Yu et al., 2009). Simply, the present model's level of abstraction does not allow for modeling such effects. However, future development of its visual buffer component, which in the present study helped to account for the effects of temporal and spatial disintegration of stimuli, might also enable the present model to replicate some more specific attentional effects in the Stroop.

In summary, the comparison of the integrated utility-based model with its competitors revealed that it predicted the largest number of Stroop effects, being the only model that accounted for all four types of Stroop phenomena considered. At the same time, the neural networks fail to account for the practice effect, as well as the stimulus-related effects, whereas Weaver++ and the utility-learning model are unable, in their present form, to explain the performance dynamics and adaptation. Most importantly, none of the alternative models can account for the three specific response-related effects addressed in Simulations 16–18. In addition to the substantial qualitative agreement of our simulations with experimental data, their quantitative fit was also satisfactory. Figure 24 presents the scatterplot of observed and simulated data points pertaining to the response latency as well as to the latency congruency effects. Because each group of data points displays a different time scale, comparing them altogether would lead to an overestimation of the correlation between the observed and simulated data. In order to bring the two groups to one and the same scale, data points were normalized separately in each group, by subtracting the group mean from each data point, and dividing the result by the group standard deviation. Then we estimated the correlation using the mixture of the normalized data. The Pearson correlation between predictions and observations equaled $r = .970$, $p < .001$, meaning that 94.1% of observed variance was successfully explained by the model. The mean error of prediction equaled $RMSD = 14$ ms for congruency effect and $RMSD = 31$ ms for response latency, which was reasonably low. The fit of 100 simulated data points was obtained optimizing only nine original parameters, plus 15 parameter values additionally optimized for Simulations 4, 5, 11, 12, 14, and 17 (in total, 24 parameters were optimized). Moreover, beyond the latency data, with no further parameter modification, the model aptly predicted the response accuracy (Simulation 1), the distributions of response latency and accuracy as well as of the congruency effect (Simulations 2 and 3), the negative correlation between incongruent trials accuracy and the latency congruency effect (Simulation 3), as well as the overall shapes of the N2/N450 and ERN waves (Simulations 9 and 10) and the specific effect concerning the N450 wave (Simulation 18).

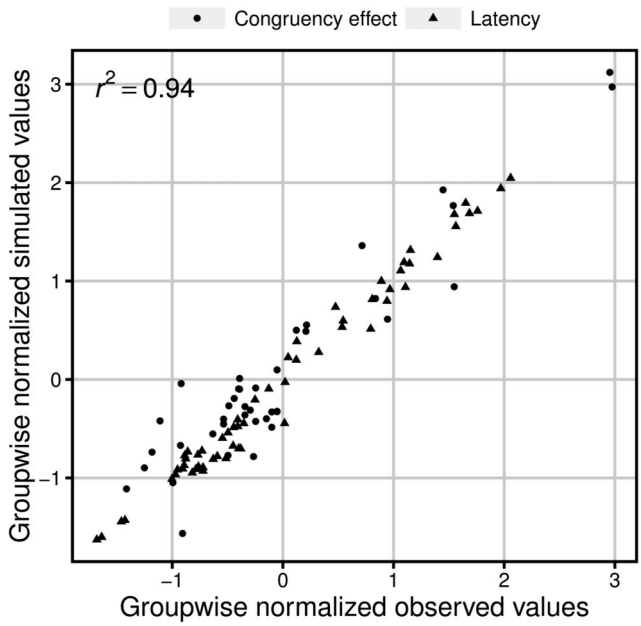

Figure 24. Scatter plot of (normalized) response latency and congruency effects observed in respective empirical data and simulated by the integrated utility-based model. For description of the normalization procedure see text.

Theoretical Contribution

The integrated utility-based model explains the resolution of Stroop interference using three crucial control mechanisms: (a) response utility learning, (b) response conflict evaluation, and (c) the conflict-driven top-down control promoting the selection of goal-relevant over goal-irrelevant responses. Each mechanism has proven to be highly useful in accounting for executive control phenomena (including Stroop). The present study demonstrated that, when applied together, the three mechanisms suffice to explain most of the known Stroop effects. Importantly, these mechanisms were not just merged into a kind of "Swiss-army knife." A key theoretical contribution of the present model consists of the close integration of the utility-based action selection with conflict evaluation and control regulation. All three become critically entangled in the "evaluate utility-based conflict—adapt control—select the most utile rule—reinforce its utility" cycle. There are three most important points of such an integration.

First, unlike the neural networks, which evaluate conflict among response node activations, the integrated utility-based model computes the current conflict level using utilities of competing responses. Moreover, whether two rules are judged competing (or not) depends on whether they yield either compatible or divergent

outcomes. Overall, this means that what matters for conflict evaluation is how probable, in a given situation, the selection of alternative actions is (i.e., how likely they are in that very task and under that strength of control) in comparison with the most goal-relevant action. In consequence, this novel approach integrates two to date quite isolated (but see Shenhav, Botvinick, & Cohen, 2013) lines of modeling: utility-based action selection and conflict evaluation.

Second, conflict-driven top-down control affects the momentary utility of competing actions, not—like in the neural networks—just their sheer activation. Strong control substantially decreases utility of goal-irrelevant actions, even if overall they are highly utile (i.e., they have been helpful in many situations, but not in the current one), helping to select weaker but more goal-relevant actions. As a result, the model directly implements the proposed notion of control, whose function is to alter the probability distribution of actions from the one primarily determined by the learned effectiveness of actions (their base utilities) into the one primarily determined by the goal-relevance of actions (their momentary utilities, related to rule-goal associations). However, achieving the latter distribution to a full extent would require so much effort (control strength) that, in fact, only some mixture distribution can be fulfilled.

Third, reinforcement learning of action utility itself depends on effective conflict evaluation and the subsequent adaptation of the control strength to the conflict level. Under high conflict, the model more quickly learns the utilities of goal-relevant actions if control is strong because it more often selects the correct, though not the most utile, rule (and boosts its base utility). When control is weak and the model often selects incorrect rules, in effect either committing errors or responding correctly but due to the wrong reason (so not boosting the correct rule's base utility), utility learning becomes less effective.

The close integration of these three cognitive control mechanisms allowed our model to replicate a broad range of Stroop phenomena, which so far has not been predicted by one and the same model. Some of phenomena required just one control mechanism, but most of them relied on more mechanisms. For example, as displayed in Figure 25, when the conflict-to-control adaptation was turned off (parameter $a = 0$) and a moderate strength of control was exerted in a constant way, the model was unable to replicate the size of the observed Gratton effect (originally consisting of 33% drop in the congruency effect after incongruent, compared with congruent, trials) by using only utility learning (it yielded an incorrect 3% increase in the effect). When utility learning was turned off but adaptation was turned on, the model predicted a minimal effect (6% drop). Only when both mechanisms were active, the observed and predicted (23% drop) effect sizes were comparable. The correct replication of the observed size of the proportion-congruent effect (59% drop in the congruency effect from the 75%- to 25%-congruent sequences) was sufficiently approximated using only adaptation (37% drop). When the latter was switched off but the utility learning was turned on, the model no longer fitted (there was an incorrect 5% increase in the effect). However, using both mechanisms slightly increased the fit (to 41% drop). The observed ISCP effect size (41% drop in the congruency effect from the 75%- to 25%-congruent stimuli) was best matched when using both the control-to-conflict adaptation and the utility learning mechanisms (25% drop), in comparison

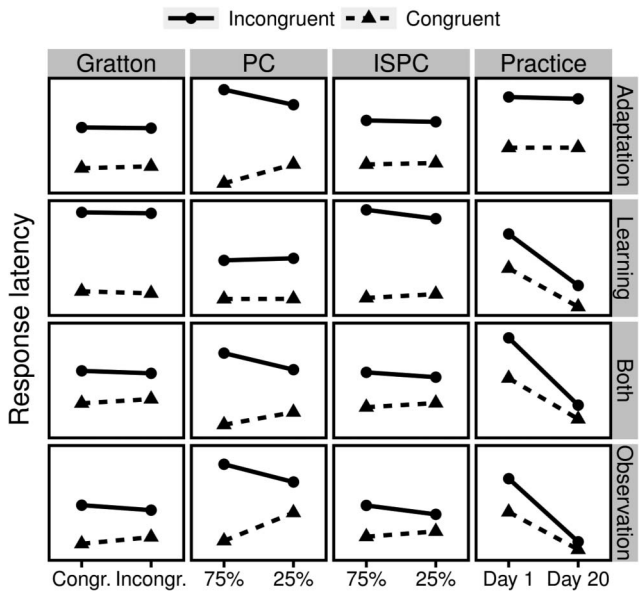

Figure 25. The Gratton effect, the regular proportion-congruent effect (PC), the item-specific proportion-congruent effect (ISPC), and the practice effect simulated by variants of the model in which: control-to-conflict adaptation was switched on but utility reinforcement learning was switched off (adaptation), adaptation was switched off but reinforcement learning was switched on (learning), and (both) adaptation and learning were switched on. For comparison, the observed data were depicted in the bottom row. For descriptions of effects see sections on Simulations 5–8.

with when the model relied solely on either the former (6% drop) or the latter mechanism (14% drop). The correct replication of the practice effect required both utility learning and adaptation, as the precisely predicted huge drop (65%) in congruency effect for shape naming from Day 1 to Day 20 was visibly attenuated when adaptation had been turned off (the model yielded only 40% drop). Of course, with sole adaptation (i.e., no learning), virtually no drop (2%) in congruency effect was noticed.

Another theoretical contribution of the integrated utility-based model consists of the generalization of control mechanisms onto situations that often occur in real life, but to date have not been examined within the Stroop task, in which more than one action can be applied correctly in response to a particular stimulus. In order to properly account for multiresponse variant of the Stroop task, the Hopfield energy used to date in the conflict monitoring theory (Botvinick et al., 2001) was substituted with the Festinger (1957) formula. Although the latter proposal has widely been applied in social psychology, it has never been adopted in the context of Stroop modeling. Here, it proved itself crucial for the generation of the null response-set size effect, as well as for the prediction of differences in EEG indices of conflict between trials in which the number of applicable rules varied between the targets and distractors (Simulations 17 and 18).

The integrated utility-based model, which lacked any semantic memory/language production mechanisms, but nevertheless accounted for the largest set of Stroop phenomena thus far, strongly implicates that the locus of conflict/interference in the Stroop task resides primarily at the response selection stage. Although memory access or linguistic processes most likely invoke additional inter-

ference when coping with the Stroop stimuli, our model suggests that the core of each Stroop effect (even the word-word interference and the semantic gradient effect) is rooted in factors that influence response selection, and can be replicated without adhering to the memory or/and language systems. Although this thesis has long been postulated by some researchers (e.g., Kornblum et al., 1990), it has been relatively less popular in Stroop modeling than are the alternative accounts referring to conflicts at the semantic/utterance level (e.g., Altmann & Davidson, 2001; Juvina & Taatgen, 2009; Roelofs, 2003).

Another influential account of cognitive control, both in general terms (Posner & Snyder, 1975; Shiffrin & Schneider, 1977), as well as more directly applied to the Stroop task (Hunt & Lansman, 1986; Kornblum et al., 1990; Roelofs, 2003), consists of dual-route models of control. According to this account, responses can be selected due to two qualitatively different processing routes. A controlled (or deliberate) route translates the relevant feature(s) of a stimulus into the response using arbitrary mapping defined by task instructions (e.g., a task set encoded in working memory). This flexible route co-occurs with an automatic (or direct) route that directly activates responses using those stimulus features that overlap with response features along some dimension. In contrast, the integrated utility-based model rejects qualitative differences between conflicting processes, and—like the neural networks and the Lovett (2005) model—postulates only the quantitative differences (here related to response utility). Under no control in the model, the dominant rules chosen due to their maximum base utility can be treated as relatively "automatic," but still under such circumstances, the selection of nondominant rules is possible (though less probable), and the latter do interfere with the former (as evidenced by the nonzero reverse interference effect)—the two outcomes to be rejected by the dual-route models. It seems that the dual-route approach may be unable to parsimoniously explain all types of dynamic interactions between dominant and nondominant processes, reflected by such phenomena as the practice, proportion-congruent, ISPC, and Gratton effects (Simulations 5–8). Of course, we do not attempt to deny the widely observed structural changes in processing paths (e.g., reading) due to long-term practice (Green & Bavelier, 2008), but we hold that one does not need to account for these changes to explain the Stroop phenomena.

Finally, an important distinction related to the control of actions consists of the difference between model-free (direct) and model-based (indirect) reinforcement learning (Dayan & Daw, 2008; Sutton & Barto, 1998). In both of these approaches, RL is aimed at the improvement of a behavioral policy in order to maximize an aggregated value of the to-be-received rewards. In model-free RL, the selection of actions is made solely on the basis of previous rewards (discounted in some way) that were associated with each action and were encoded in episodic memory. In contrast, in model-based RL, the agent attempts to learn a model of its environment, and then by using that model, it tries to predict the consequences of actions before they are taken, allowing the agent to simulate rewards in the absence of real actions. In consequence, model-based RL integrates both learning on the basis of past experience and predicting the value of future actions. In the integrated utility-based model, utility learning is an instance of model-free RL, as the sole information it uses is the feedback history. Model-based RL could be implemented in the present model in a simple way, that is, if it started to learn goal associations (values *A*) throughout its experience. Under increased control, those learned associations would affect the distribution of the expected utility of rules (thus, also their rewards) in the particular task context, constituting a model of the environment. However, in the current simulations, a predefined (though theoretically justified) set of goal association values was adopted, and the extension of the model's learning mechanisms onto model-based RL would be beyond the scope of the present study.

Conclusion

We have proposed the hybrid, symbolic-subsymbolic computational model of the three integrated mechanisms underlying effective cognitive control in the Stroop task: response utility learning, utility-based conflict evaluation, and the conflict-driven top-down control promoting the selection of goal-relevant over goal-irrelevant responses. Their complex interaction led to the replication of the largest set of Stroop phenomena thus far. In particular, the basic congruency, performance dynamics/adaptation, stimulus-related, and response-related effects have been explained. Moreover, the model substantially extended the concept of conflict evaluation, first, by introducing the well-known (but to date absent in cognitive modeling) Festinger formula of conflict, which can account for realistic cases when more than one correct response can be elicited in a particular situation, and second, by defining conflict evaluation at the level of response utility, not just response activation or strength. Also, the model explicitly implemented the notion of cognitive control as the change in probability distribution of the response options, from one primarily determined by the learned effectiveness of options into one mainly determined by their goal-relevance. Finally, the model suggests that the core part of Stroop interference is primarily rooted in response selection, and can be effectively explained without accounting for conflicts within semantic memory or language production system. In general, the model seems to substantially enrich our understanding of the way in which the mind/brain internally controls its own functioning, suggesting that effective cognitive control does not rely on a unitary control system, but emerges from the dynamic interplay of various control mechanisms specialized in the learning, evaluation, and selection of information.

References

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience, 14,* 1338–1344. http://dx.doi.org/10.1038/nn.2921

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: Laurence Erlbaum.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Cambridge, MA: MIT Press.

Austin, J. T., & Vancouver, J. B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin, 120,* 338–375. http://dx.doi.org/10.1037/0033-2909.120.3.338

Berlyne, D. E. (1960). *Conflict, arousal, and curiosity*. New York, NY: McGraw-Hill. http://dx.doi.org/10.1037/11164-000

Blais, C., & Besner, D. (2006). Reverse Stroop effects with untranslated responses. *Journal of Experimental Psychology: Human Perception and*

Performance, 32, 1345–1353. http://dx.doi.org/10.1037/0096-1523.32.6.1345

Blais, C., & Besner, D. (2007). A reverse Stroop effect without translation or reading difficulty. *Psychonomic Bulletin & Review, 14,* 466–469. http://dx.doi.org/10.3758/BF03194090

Blais, C., Robidoux, S., Risko, E. F., & Besner, D. (2007). Item-specific adaptation and the conflict-monitoring hypothesis: A computational model. *Psychological Review, 114,* 1076–1086. http://dx.doi.org/10.1037/0033-295X.114.4.1076

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. http://dx.doi.org/10.1037/0033-295X.108.3.624

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences, 16,* 106–113. http://dx.doi.org/10.1016/j.tics.2011.12.010

Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science, 307,* 1118–1121.

Brown, J. W., & Braver, T. S. (2008). A computational model of risk, conflict and individual difference effects in the anterior cingulate cortex. *Brain Research, 1202,* 99–108.

Chuderski, A., Senderecka, M., Kałamała, P., Kroczek, B., & Ociepka, M. (2015). *When some roads lead to Rome but others lead elsewhere: ERP correlates of the conflict level in the multi-response Stroop task*. Unpublished manuscript.

Chuderski, A., Smoleń, T., & Taraday, M. (2014). Neither a response nor stimulus set-size effect in the manual Stroop task. *Studia Psychologica, 56,* 21–35.

Cockburn, J., & Frank, M. J. (2011). Reinforcement learning, conflict monitoring, and cognitive control: An integrative model of cingulate-striatal interactions and the ERN. In R. B. Mars, J. Sallet, M. F. S. Rushworth, & N. Yeung (Eds.), *Neural basis of motivational and cognitive control* (pp. 311–331). Cambridge, MA: MIT Press. http://dx.doi.org/10.7551/mitpress/9780262016438.003.0017

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. http://dx.doi.org/10.1037/0033-295X.97.3.332

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

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review, 99,* 45–77. http://dx.doi.org/10.1037/0033-295X.99.1.45

Cohen, J. D., Usher, M., & McClelland, J. L. (1998). A PDP approach to set size effects within the Stroop task: Reply to Kanne, Balota, Spieler, and Faust (1998). *Psychological Review, 105,* 188–194. http://dx.doi.org/10.1037/0033-295X.105.1.188

Crump, M. J. C., Gong, Z., & Milliken, B. (2006). The context-specific proportion congruent Stroop effect: Location as a contextual cue. *Psychonomic Bulletin & Review, 13,* 316–321. http://dx.doi.org/10.3758/BF03193850

Dallas, M., & Merikle, E. M. (1976). Semantic processing of non-attended visual information. *Canadian Journal of Psychology, 30,* 15–21. http://dx.doi.org/10.1037/h0082040

Dalrymple-Alford, E. C. (1972). Associative facilitation and interference in the Stroop color-word task. *Perception and Psychophysics, 11,* 274–276.

Dalrymple-Alford, E. C., & Budayr, B. (1966). Examination of some aspects of the Stroop Color-Word Test. *Perceptual and Motor Skills, 23,* 1211–1214. http://dx.doi.org/10.2466/pms.1966.23.3f.1211

Davelaar, E. J. (2008). A computational study of conflict-monitoring at two levels of processing: Reaction time distributional analyses and hemodynamic responses. *Brain Research, 1202,* 109–119. http://dx.doi.org/10.1016/j.brainres.2007.06.068

Davelaar, E. J., & Stevens, J. (2009). Sequential dependencies in the Eriksen flanker task: A direct comparison of two competing accounts. *Psychonomic Bulletin & Review, 16,* 121–126. http://dx.doi.org/10.3758/PBR.16.1.121

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective & Behavioral Neuroscience, 8,* 429–453. http://dx.doi.org/10.3758/CABN.8.4.429

Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences of the United States of America, 95,* 14529–14534. http://dx.doi.org/10.1073/pnas.95.24.14529

De Pisapia, N., Repovs, G., & Braver, T. S. (2008). Computational models of attention and cognitive control. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 422–450). Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511816772.019

Dunbar, K., & MacLeod, C. M. (1984). A horse race of a different color: Stroop interference patterns with transformed words. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 622–639. http://dx.doi.org/10.1037/0096-1523.10.5.622

Dyer, F. N. (1971). The duration of word meaning responses: Stroop interference for different preexposures of the word. *Psychonomic Science, 25,* 229–231. http://dx.doi.org/10.3758/BF03329102

Dyer, F. N. (1973). Interference and facilitation for color naming with separate bilateral presentations of the word and color. *Journal of Experimental Psychology, 99,* 314–317. http://dx.doi.org/10.1037/h0035245

Dyer, F. N., & Severance, L. J. (1972). Effects of irrelevant colors on reading of color names: A control replication of the "reversed Stroop" effect. *Psychonomic Science, 28,* 336–338. http://dx.doi.org/10.3758/BF03328756

Dyer, F. N., & Severance, L. J. (1973). Stroop interference with successive presentations of separate incongruent words and colors. *Journal of Experimental Psychology, 98,* 438–439. http://dx.doi.org/10.1037/h0034353

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16,* 143–149. http://dx.doi.org/10.3758/BF03203267

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology, 58,* 203–210. http://dx.doi.org/10.1037/h0041593

Garner, W. R. (1962). *Uncertainty and structure as psychological concepts*. New York, NY: Wiley.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

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. http://dx.doi.org/10.1037/0096-1523.8.6.875

Gollwitzer, P. M. (1993). Goal achievement: The role of intentions. *European Review of Social Psychology, 4,* 141–185. http://dx.doi.org/10.1080/14792779343000059

Gratton, G., Coles, M. G. H., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology: General, 121,* 480–506. http://dx.doi.org/10.1037/0096-3445.121.4.480

Green, C. S., & Bavelier, D. (2008). Exercising your brain: A review of human brain plasticity and training-induced learning. *Psychology and Aging, 23,* 692–701. http://dx.doi.org/10.1037/a0014345

Grinband, J., Savitskaya, J., Wager, T. D., Teichert, T., Ferrera, V. P., & Hirsch, J. (2011). The dorsal medial frontal cortex is sensitive to time on task, not response conflict or error likelihood. *NeuroImage, 57,* 303–311. http://dx.doi.org/10.1016/j.neuroimage.2010.12.027

Gruber, O., & Goschke, T. (2004). Executive control emerging from dynamic interactions between brain systems mediating language, working memory and attentional processes. *Acta Psychologica, 115,* 105–121. http://dx.doi.org/10.1016/j.actpsy.2003.12.003

Hasher, L., Lustig, C., & Zacks, R. T. (2007). Inhibitory mechanisms and the control of attention. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 227–249). New York, NY: Oxford University Press.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Towards an executive without a homunculus: Computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 362,* 1601–1613. http://dx.doi.org/10.1098/rstb.2007.2055

Hentschel, U. (1973). Two new interference tests compared to the Stroop color-word test. *Psychological Research Bulletin, Lund University, 13,* 1–24.

Herd, S. A., Banich, M. T., & O'Reilly, R. C. (2006). Neural mechanisms of cognitive control: An integrative model of Stroop task performance and FMRI data. *Journal of Cognitive Neuroscience, 18,* 22–32. http://dx.doi.org/10.1162/089892906775250012

Hohnsbein, J., Falkenstein, M., & Hoorman, J. (1989). Error processing in visual and auditory choice reaction tasks. *Journal of Psychophysiology, 3,* 32.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109,* 679–709. http://dx.doi.org/10.1037/0033-295X.109.4.679

Holroyd, C. B., & Yeung, N. (2011). An integrative theory of anterior cingulate cortex function: Option selection in hierarchical reinforcement learning. In R. B. Mars, J. Sallet, M. F. S. Rushworth, & N. Yeung (Eds.), *Neural basis of motivational and cognitive control* (pp. 333–349). Cambridge, MA: MIT Press. http://dx.doi.org/10.7551/mitpress/9780262016438.003.0018

Holroyd, C. B., Yeung, N., Coles, M. G. H., & Cohen, J. D. (2005). A mechanism for error detection in speeded response time tasks. *Journal of Experimental Psychology: General, 134,* 163–191.

Hunt, E., & Lansman, M. (1986). Unified model of attention and problem solving. *Psychological Review, 93,* 446–461. http://dx.doi.org/10.1037/0033-295X.93.4.446

Jacoby, L. L., Lindsay, D. S., & Hessels, S. (2003). Item-specific control of automatic processes: Stroop process dissociations. *Psychonomic Bulletin & Review, 10,* 638–644. http://dx.doi.org/10.3758/BF03196526

Jacoby, L. L., McElree, B., & Trainham, T. N. (1999). Automatic influences as accessibility bias in memory and Stroop-like tasks: Toward a formal model. In A. Koriat & D. Gopher (Eds.), *Attention and performance XVII* (pp. 461–486). Cambridge, MA: MIT Press.

Jensen, A. R., & Rohwer, W. D., Jr. (1966). The Stroop color-word test: A review. *Acta Psychologica, 25,* 36–93. http://dx.doi.org/10.1016/0001-6918(66)90004-7

Jiang, J., Heller, K., & Egner, T. (2014). Bayesian modeling of flexible cognitive control. *Neuroscience and Biobehavioral Reviews, 46,* 30–43. http://dx.doi.org/10.1016/j.neubiorev.2014.06.001

Jones, A. D., Cho, R. Y., Nystrom, L. E., Cohen, J. D., & Braver, T. S. (2002). A computational model of anterior cingulate function in speeded response tasks: Effects of frequency, sequence, and conflict. *Cognitive, Affective & Behavioral Neuroscience, 2,* 300–317. http://dx.doi.org/10.3758/CABN.2.4.300

Juvina, I., & Taatgen, N. A. (2009). A repetition-suppression account of between-trial effects in a modified Stroop paradigm. *Acta Psychologica, 131,* 72–84. http://dx.doi.org/10.1016/j.actpsy.2009.03.002

Kahneman, D., & Henik, A. (1981). Perceptual organization and attention. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 181–211). Hillsdale, NJ: Erlbaum.

Kanne, S. M., Balota, D. A., Spieler, D. H., & Faust, M. E. (1998). Explorations of Cohen, Dunbar, and McClelland's (1990) connectionist model of Stroop performance. *Psychological Review, 105,* 174–187. http://dx.doi.org/10.1037/0033-295X.105.1.174

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. http://dx.doi.org/10.1126/science.1089910

Kimberg, D. Y., & Farah, M. J. (1993). A unified account of cognitive impairments following frontal lobe damage: The role of working memory in complex, organized behavior. *Journal of Experimental Psychology: General, 122,* 411–428. http://dx.doi.org/10.1037/0096-3445.122.4.411

Klein, G. S. (1964). Semantic power measured through the interference of words with color-naming. *The American Journal of Psychology, 77,* 576–588. http://dx.doi.org/10.2307/1420768

Kopp, B., Rist, F., & Mattler, U. (1996). N200 in the flanker task as a neurobehavioral tool for investigating executive control. *Psychophysiology, 33,* 282–294. http://dx.doi.org/10.1111/j.1469-8986.1996.tb00425.x

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility—A model and taxonomy. *Psychological Review, 97,* 253–270. http://dx.doi.org/10.1037/0033-295X.97.2.253

La Heij, W., & van den Hof, E. (1995). Picture-word interference increases with target-set size. *Psychological Research, 58,* 119–133. http://dx.doi.org/10.1007/BF00571100

Larson, M. J., Clayson, P. E., & Clawson, A. (2014). Making sense of all the conflict: A theoretical review and critique of conflict-related ERPs. *International Journal of Psychophysiology, 93,* 283–297. http://dx.doi.org/10.1016/j.ijpsycho.2014.06.007

Larson, M. J., Kaufman, D. A., & Perlstein, W. M. (2009). Neural time course of conflict adaptation effects on the Stroop task. *Neuropsychologia, 47,* 663–670. http://dx.doi.org/10.1016/j.neuropsychologia.2008.11.013

Lewin, K. (1935). *A dynamic theory of personality*. New York, NY: McGraw-Hill.

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. http://dx.doi.org/10.1037/0096-1523.20.2.219

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 & Cognition, 7,* 166–174. http://dx.doi.org/10.3758/BF03197535

Lovett, M. C. (2001). Selective attention and strategies: Not just another model of Stroop. In E. M. Altmann, A. Cleeremans, & C. D. Schunn (Eds.), *Proceedings of the Fourth International Conference on Cognitive Modeling* (pp. 151–156). Mahwah, NJ: Erlbaum.

Lovett, M. C. (2005). A strategy-based interpretation of Stroop. *Cognitive Science, 29,* 493–524. http://dx.doi.org/10.1207/s15516709cog0000_24

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: Wiley.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109,* 163–203. http://dx.doi.org/10.1037/0033-2909.109.2.163

MacLeod, C. M. (1998). Training on integrated versus separated Stroop tasks: The progression of interference and facilitation. *Memory & Cognition, 26,* 201–211. http://dx.doi.org/10.3758/BF03201133

MacLeod, C. M., Chiappe, D. L., & Fox, E. F. (2002). The crucial roles of stimulus matching and stimulus identity in negative priming. *Psycho-*

nomic Bulletin & Review, 9, 521–528. http://dx.doi.org/10.3758/BF03196308

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, 14,* 126–135. http://dx.doi.org/10.1037/0278-7393.14.1.126

Mayr, U., Awh, E., & Laurey, P. (2003). Does conflict adaptation require executive control? *Nature Neuroscience, 6,* 450–452.

McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.

Melara, R. D., & Algom, D. (2003). Driven by information: A tectonic theory of Stroop effects. *Psychological Review, 110,* 422–471. http://dx.doi.org/10.1037/0033-295X.110.3.422

Mewhort, D. J., Braun, J. G., & Heathcote, A. (1992). Response time distributions and the Stroop Task: A test of the Cohen, Dunbar, and McClelland (1990) model. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 872–882. http://dx.doi.org/10.1037/0096-1523.18.3.872

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York, NY: Henry Holt. http://dx.doi.org/10.1037/10039-000

Monsell, S., & Driver, J. (2000). Banishing the control homunculus. In J. Driver & S. Monsell (Eds.), *Control of cognitive processes, attention and performance XVIII* (pp. 3–31). Cambridge, MA: MIT Press.

Munakata, Y., Herd, S. A., Chatham, C. H., Depue, B. E., Banich, M. T., & O'Reilly, R. C. (2011). A unified framework for inhibitory control. *Trends in Cognitive Sciences, 15,* 453–459. http://dx.doi.org/10.1016/j.tics.2011.07.011

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology, 9,* 353–383. http://dx.doi.org/10.1016/0010-0285(77)90012-3

Neal, R. M. (2003). Slice sampling. *Annals of Statistics, 31,* 705–767. http://dx.doi.org/10.1214/aos/1056562461

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Nieuwenhuis, S., Schweizer, T. S., Mars, R. B., Botvinick, M. M., & Hajcak, G. (2007). Error-likelihood prediction in the medial frontal cortex: A critical evaluation. *Cerebral Cortex, 17,* 1570–1581. http://dx.doi.org/10.1093/cercor/bhl068

Nigg, J. T. (2000). On inhibition/disinhibition in developmental psychology: Views from cognitive and personality psychology and a working inhibition taxonomy. *Psychological Bulletin, 126,* 220–246.

Norman, D. A. (1981). Categorization of action slips. *Psychological Review, 88,* 1–15. http://dx.doi.org/10.1037/0033-295X.88.1.1

Norman, D. A., & Shallice, T. (1986). Attention and action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation: Advances in research and theory* (Vol. 4, pp. 1–18). New York, NY: Plenum Press. http://dx.doi.org/10.1007/978-1-4757-0629-1_1

Phaf, R. H., Van der Heijden, A. H. C., & Hudson, P. T. W. (1990). SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology, 22,* 273–341. http://dx.doi.org/10.1016/0010-0285(90)90006-P

Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 55–85). Hillsdale, NJ: Erlbaum.

Proctor, R. W. (1978). Sources of color-word interference in the Stroop color-naming task. *Perception & Psychophysics, 23,* 413–419. http://dx.doi.org/10.3758/BF03204145

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin, 86,* 446–461. http://dx.doi.org/10.1037/0033-2909.86.3.446

Reason, J. (1990). *Human error*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139062367

Roelofs, A. (2000). *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. (2001). Set size and repetition matter: Comment on Caramazza and Costa (2000). *Cognition, 80,* 283–290. http://dx.doi.org/10.1016/S0010-0277(01)00134-2

Roelofs, A. (2003). Goal-referenced selection of verbal action: Modeling attentional control in the Stroop task. *Psychological Review, 110,* 88–125. http://dx.doi.org/10.1037/0033-295X.110.1.88

Roelofs, A. (2012). Attention, spatial integration, and the tail of response time distributions in Stroop task performance. *Quarterly Journal of Experimental Psychology, 65,* 135–150. http://dx.doi.org/10.1080/17470218.2011.605152

Roelofs, A., & Lamers, M. (2007). Modelling the control of visual attention in Stroop-like tasks. In A. S. Meyer, L. R. Wheeldon, & A. Krott (Eds.), *Automacity and control in language processing* (pp. 123–142). Hove, UK: Psychology Press.

Scheffers, M. K., & Coles, M. G. H. (2000). Performance monitoring in a confusing world: Error-related brain activity, judgments of response accuracy, and types of errors. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 141–151. http://dx.doi.org/10.1037/0096-1523.26.1.141

Scherbaum, S., Dshemuchadse, M., Ruge, H., & Goschke, T. (2012). Dynamic goal states: Adjusting cognitive control without conflict monitoring. *NeuroImage, 63,* 126–136. http://dx.doi.org/10.1016/j.neuroimage.2012.06.021

Schmidt, J. R. (2013). The Parallel Episodic Processing (PEP) model: Dissociating contingency and conflict adaptation in the item-specific proportion congruent paradigm. *Acta Psychologica, 142,* 119–126. http://dx.doi.org/10.1016/j.actpsy.2012.11.004

Seymour, E. H. (1973). Stroop interference in naming and verifying spatial locations. *Perception & Psychophysics, 14,* 95–100. http://dx.doi.org/10.3758/BF03198622

Shallice, T., & Burgess, P. W. (1993). Supervisory control of action and thought selection. In A. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness, and control* (pp. 171–181). Oxford, UK: Clarendon Press.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron, 79,* 217–240.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84,* 127–190. http://dx.doi.org/10.1037/0033-295X.84.2.127

Shimada, H., & Nakajima, Y. (1991). Double response to Stroop stimuli. *Perceptual and Motor Skills, 73,* 571–574. http://dx.doi.org/10.2466/pms.1991.73.2.571

Shor, R. E. (1971). Symbol processing speed differences and symbol interference effects in a variety of concept domains. *Journal of General Psychology, 85,* 187–205. http://dx.doi.org/10.1080/00221309.1971.9920672

Spieler, D. H., Balota, D. A., & Faust, M. E. (2000). Levels of selective attention revealed through analyses of response time distributions. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 506–526. http://dx.doi.org/10.1037/0096-1523.26.2.506

Stafford, T., & Gurney, K. N. (2007). Biologically constrained action selection improves cognitive control in a model of the Stroop task. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 362,* 1671–1684. http://dx.doi.org/10.1098/rstb.2007.2060

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643–662. http://dx.doi.org/10.1037/h0054651

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Taatgen, N. A. (2007). The minimal control principle. In W. Gray (Ed.), *Integrated models of cognitive systems* (pp. 368–379). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195189193.003.0025

Taatgen, N. A., & Lee, F. J. (2003). Production compilation: A simple mechanism to model complex skill acquisition. *Human Factors, 45,* 61–76. http://dx.doi.org/10.1518/hfes.45.1.61.27224

Thurstone, L. L. (1944). *A factorial study of perception*. Chicago, IL: University of Chicago Press.

Tzelgov, J., Henik, A., & Berger, J. (1992). Controlling Stroop effects by manipulating expectations for color words. *Memory & Cognition, 20,* 727–735. http://dx.doi.org/10.3758/BF03202722

van Maanen, L., & Van Rijn, H. (2007). An accumulator model of semantic interference. *Cognitive Systems Research, 8,* 174–181. http://dx.doi.org/10.1016/j.cogsys.2007.05.002

van Maanen, L., van Rijn, H., & Borst, J. P. (2009). Stroop and picture-word interference are two sides of the same coin. *Psychonomic Bulletin & Review, 16,* 987–999. http://dx.doi.org/10.3758/PBR.16.6.987

Van Veen, V., & Carter, C. S. (2002). The timing of action-monitoring processes in the anterior cingulate cortex. *Journal of Cognitive Neuroscience, 14,* 593–602. http://dx.doi.org/10.1162/08989290260045837

Verguts, T., & Notebaert, W. (2008). Hebbian learning of cognitive control: Dealing with specific and nonspecific adaptation. *Psychological Review, 115,* 518–525. http://dx.doi.org/10.1037/0033-295X.115.2.518

West, R., & Alain, C. (2000). Effects of task context and fluctuations of attention on neural activity supporting performance of the Stroop task. *Brain Research, 873,* 102–111. http://dx.doi.org/10.1016/S0006-8993(00)02530-0

West, R., Bailey, K., Tiernan, B. N., Boonsuk, W., & Gilbert, S. (2012). The temporal dynamics of medial and lateral frontal neural activity related to proactive cognitive control. *Neuropsychologia, 50,* 3450–3460. http://dx.doi.org/10.1016/j.neuropsychologia.2012.10.011

White, B. W. (1969). Interference in identifying attributes and attribute names. *Perception & Psychophysics, 6,* 166–168. http://dx.doi.org/10.3758/BF03210086

White, C. N., Ratcliff, R., & Starns, J. J. (2011). Diffusion models of the flanker task: Discrete versus gradual attentional selection. *Cognitive Psychology, 63,* 210–238. http://dx.doi.org/10.1016/j.cogpsych.2011.08.001

Yeung, N. (2012). Conflict monitoring and cognitive control. In K. N. Ochsner & S. M. Kosslyn (Eds.), *Oxford handbook of cognitive neuroscience* (pp. 275–299). Oxford, UK: Oxford University Press.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review, 111,* 931–959. http://dx.doi.org/10.1037/0033-295X.111.4.931

Yeung, N., & Cohen, J. D. (2006). The impact of cognitive deficits on conflict monitoring. Predictable dissociations between the error-related negativity and N2. *Psychological Science, 17,* 164–171. http://dx.doi.org/10.1111/j.1467-9280.2006.01680.x

Yeung, N., Cohen, J. D., & Botvinick, M. M. (2011). Errors of interpretation and modeling: A reply to Grinband et al. *NeuroImage, 57,* 316–319. http://dx.doi.org/10.1016/j.neuroimage.2011.04.029

Yeung, N., & Nieuwenhuis, S. (2009). Dissociating response conflict and error likelihood in anterior cingulate cortex. *The Journal of Neuroscience, 29,* 14506–14510. http://dx.doi.org/10.1523/JNEUROSCI.3615-09.2009

Yu, A. J., Dayan, P., & Cohen, J. D. (2009). Dynamics of attentional selection under conflict: Toward a rational Bayesian account. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 700–717. http://dx.doi.org/10.1037/a0013553

Appendix

Experiments

Experiment 1

This experiment collected a large dataset on individual differences in latency in both the color-word and the picture-word tasks. The study also confirmed that the priming effects, commonly observed in the vocal Stroop (see Juvina & Taatgen, 2009), are absent in the manual variant.

Participants

Volunteer participants, recruited via social networking websites, were tested in groups of six to 12 people. Each participant gave informed consent and was paid the equivalent of 10 euro in Polish zloty. A total of 210 women and 138 men participated (M age = 24.3 years, range 17 – 45).

Materials and Procedure

Two tasks were applied in a random order. The stimuli in the color-word task were four capital Polish words (approximately 7 cm × 2 cm in size), printed in bold Arial font, naming the colors red, green, blue, and brown. Each word could be displayed in any of the four ink colors. In congruent stimuli (random 50 trials) the word and color matched. For incongruent stimuli (another 50 trials), they differed. The stimuli in the figure-word task (Hentschel, 1973) consisted of four figures (square, rhombus, circle, and trapezium, all approximately 4 cm in size), printed in blue with a black border. Each congruent stimulus (random 50 trials) contained a word (approximately 4 cm × 1 cm in size, printed in regular Arial font) for this very figure. For incongruent stimuli (another 50 trials), a word denoted another figure. The stimuli were presented at the center of the computer screen.

The stimuli sequence was always random, with two constraints. First, direct repetitions of the target stimuli were forbidden, in order to eliminate the exact stimulus-response repetitions (see Mayr, Awh, & Laurey, 2003). Second, the control of negative priming effects (see Juvina & Taatgen, 2009) was exerted in the following way: in 50% of incongruent trials (the primed trials), a target color matched a word presented in a preceding (i.e., N-1th) trial, whereas in another 50% of incongruent trials (the unprimed trials) the Nth target did not match a word at the N-1 position.

Trials lasted until a response was given, or for a maximum of 2.2 s. An 800 ms mask separated the subsequent trials. The instructions for both tasks were to avoid reading a word and to press a response key ("Z," "X," "N," or "M") that was assigned to a presented color/shape. Each task was preceded by a training session of 40 neutral stimuli (either "HHHHHHH" strings or figures with no word inside, respectively), to strengthen the stimulus-key associations. The proportion and latency of correct responses were measured. As we were interested in the complete RT distributions as well as in individual differences in the congruency effects, we did not trim the latencies, except for excluding reactions shorter than 250 ms (less than 0.1% of data). The independent variables were: the congruent versus incongruent trials, the color-word versus the figure-word task, and the primed versus unprimed incongruent trials.

Results and Discussion

One participant missed the figure-word task, and two participants missed the color-word task, so their data in the remaining tasks were excluded. Moreover five participants had zero accuracy in the incongruent condition (indicating that they permanently responded to words), so their results were also excluded, leaving a final sample of 340 participants. The congruency effects in both the color-word task (M_{con} = 887 ms vs. M_{inc} = 1016 ms) and in the figure word task (M_{con} = 849 ms vs. M_{inc} = 917 ms) were significant, $F(1, 339) = 536.50$, $p < .001$, $\eta^2 = .61$, $F(1, 339) = 241.0$, $p < .001$, $\eta^2 = .42$, respectively, however, the former effect was significantly larger compared with the latter effect (ΔM_{col} = 129 vs. ΔM_{fig} = 70), $F(1, 339) = 89.14$, $p < .001$, $\eta^2 = .20$. Thus, the color-word variant seemed to impose a larger load on the interference resolution mechanisms than did the figure-word task. No effect of priming was observed in the case of incongruent trials (M_{prim} = 1022 vs. M_{unprim} = 1013), $F(1, 335) = 1.38$ (four results were removed from this analysis due to zero accuracy in either the primed or unprimed conditions).

(*Appendix continues*)

Experiment 2

A decrease in the proportion of congruent trials usually leads to a decrease of the congruency effect (Logan & Zbrodoff, 1979; Tzelgov, Henik, & Berger, 1992). The conflict-monitoring model (Botvinick et al., 2001) explains this effect in terms of a permanently increased conflict/control in the primarily incongruent sequences (PIS). Crucially, Melara and Algom (2003) noted that in primarily congruent sequences (PCS), there was a stronger correlation between colors and words than in PIS. When there are no congruent trials at all, then such a correlation equals zero (i.e., each color can be associated with any distracting word), whereas when a word always matches a color, then it equals one. As people may be prone to the so-called Garner (1962, 1974) interference, consisting of the inability to filter out an irrelevant dimension of an object, which can thus yield statistical information about a (correlated) relevant dimension, the use of such information, incorrect in the case of incongruent trials, would be more frequent (and, thus, the congruency effect will be larger) in PCS than in PIS.

In order to deconfound the proportion of congruent trials and the correlation of the color and word dimensions, we used incongruent stimuli that always associated one particular value of color with one and the same word (e.g., color RED always accompanied word "blue"). Therefore, sequences containing X percent of congruent stimuli, as well as sequences containing 100 – X percent of such stimuli, yielded the same correlation (r = X) between the color and word dimensions. We aimed to observe the effect of the proportion of congruent trials on interference even if there was no difference in correlation between colors and words, and thus the explanation solely in terms of the Garner interference could be rejected.

Participants

There were 297 participants (180 females, M age = 22.6 years, range 15 – 46). Such a large sample was used because data from Stroop were used in another, correlational study. The participants were randomly assigned to two groups ("the 75%-congruent group" – 152 people; "the 25%-congruent group" – 145 people). Each condition yielded correlation of r = .75 between the color and word dimensions. Also, an additional group of 111 people (73 females, M age = 22.51 years, range 16 – 43) were tested in the 50%-congruent control condition, which yielded a weaker color-word correlation of r = .5. The same testing conditions applied as in Experiment 1.

Materials and Procedure

The color-word task was used exactly as in Experiment 1 with one exception: there were 90 congruent and 30 incongruent trials presented in the 75%-congruent group, whereas 30 congruent and 90 incongruent trials were shown in the 25%-congruent group. In each condition of the 50%-congruent Task 60 trials were shown. For incongruent stimuli, each color/figure was associated with exactly one predefined word. There was one experimental factor: the proportion of congruent stimuli. The congruency effect constituted the dependent variable.

Results and Discussion

The congruency effect was significantly higher in the 75%-congruent group than in the 25%-congruent group ($\Delta M_{75\%}$ = 196 ms vs. $\Delta M_{25\%}$ = 82 ms), $t(295) = 7.81$, $p < .001$, $d = 0.88$. The effect for the control condition was located in between ($\Delta M_{50\%}$ = 120 ms). These data suggest that the proportion of congruent trials strongly influence the congruency effect (the larger the proportion, the larger the congruency effect) even when the correlation between the color and word dimensions is controlled for. So, contrary to Melara and Algom's (2003) claim, but in line with the connectionist conflict-monitoring models (Botvinick et al., 2001), these results cannot be (solely) attributed to the Garner interference, but require some form of conflict adaptation.

Received September 4, 2014
Revision received October 29, 2015
Accepted October 30, 2015 ■