NEITHER A RESPONSE NOR STIMULUS SET-SIZE EFFECT FOUND IN THE MANUAL STROOP TASK

Adam CHUDERSKI[1], Tomasz SMOLEŃ[2], Maciej TARADAY[3]

[1]Institute of Philosophy, Jagiellonian University in Krakow
Grodzka 52, 31-044 Krakow, Poland
E-mail: adam.chuderski@gmail.com

[2]Department of Psychology, Pedagogical University of Krakow
Podchorazych 2, 30-084 Krakow, Poland

[3]Institute of Psychology, Jagiellonian University in Krakow
Mickiewicza 3, 31-120 Krakow, Poland

Abstract: Existing computational models of the Stroop task differ in predictions concerning the set-size effect, which is the relation between a number of stimuli/responses and the magnitude of the Stroop interference. However, relevant empirical data is not unequivocal, as some studies reported no set-size effects, while others found substantial set-size effects. We administered two experiments in order to resolve this discrepancy in the case of the manual Stroop task. Experiment 1 compared conditions including four, six, and eight stimulus/response mappings in the picture-word task. No reliable set-size effects were found, apart from a weak effect observed when a working memory load imposed by the task was deliberately decreased. Experiment 2 tested conditions consisting of four versus eight mappings in the color-word task, and it replicated results of Experiment 1. As both experiments had sufficient power to detect set-size effects if they existed, our data are inconsistent with models predicting such effects.

Key words: Stroop task, cognitive control, interference, set-size effects

The Stroop task is a widely used test to determine how humans control (organize, coordinate) their own cognitive processing. In its standard color-word variant, the task involves the naming of the ink color of a word that denotes the same (congruent trials) or another color (incongruent trials). The picture-word variant requires naming a picture, while ignoring an included word. The crucial observation is the *interference effect*: an increased response latency (and sometimes error rate) in the incongruent trials, compared to the congruent trials, or to trials called neutral, in which meaningless colored strings (e.g., XXXXX) are shown. This observation means that overcoming a well-learned action (i.e., reading), in order to perform a relatively novel process (naming a color or a picture), requires an additional effort from the cognitive system, and results in significant processing costs.

Research on cognitive control using the Stroop task involves showing that the interference effect can change as a result of cer-

tain manipulations constituting, for example, whether the previous trial was either congruent (larger interference) or incongruent (smaller interference), whether either congruent (larger interference) or incongruent (smaller interference) trials dominate the sequence of trials, or whether words are either semantically similar (larger interference) or dissimilar (smaller interference) to the target pictures (for a review see MacLeod, 1991). Stroop effects also increase when additional sources of interference are introduced, for example, by displaying incongruent background colors, or spewing their names with the use of a tape recorder (e.g., Daniel, 1970). They can also decrease (or even reverse) after extensive training (MacLeod, Dunbar, 1988), or when a distractor is delayed in relation to the target onset (Glaser, Glaser, 1982).

One such manipulation concerns the number of possible stimuli/responses within the Stroop task (*set size*), and it tests whether the interference effect changes as the number of colors or pictures to be named increases. However, the results regarding the set-size effect obtained so far seem to be very unclear. For example, among studies published between the sixties and the eighties, cited in the well-known review paper by MacLeod (1991), several studies showed no set size-effects (e.g., Ray, 1974), three studies yielded an increase in interference (e.g., Williams, 1977), and another three studies demonstrated its decrease (e.g., La Heij, Van der Heijden, Schreuder, 1985). MacLeod (1991, p. 184) concluded that "the manipulation of response set size seems so straightforward that these conflicting findings are frustrating." One likely reason for such frustrating results is the fact that most of those procedures were not computerized, and thus they were not sufficiently controlled (e.g.,

printed lists of colored words were used; Williams, 1977).

However, two relatively newer (and fully computerized) studies (La Heij, van den Hof, 1995; Kanne et al., 1998) did not provide any clearer view either. La Heij and van den Hof, using the picture-word task, compared conditions including 4 versus 16 pictures, and found a 40 ms increase in interference (their Experiment 1) as a function of the set size. However, 16 stimuli seem to be too large an alteration of testing conditions to allow for univocal conclusions to be drawn. For instance, working memory might be heavily loaded in that condition, as suggested by La Heij and van den Hof's 2.5 times greater error rate in the 16-stimuli compared to the 4-stimuli condition. Moreover, the complexity of the design of their experiment, which included both semantically related and unrelated distractors, could interact with the set size. It is noteworthy that in a similar study by La Heij and Vermeij (1987), using eight stimuli produced exactly the reverse pattern of data, i.e., a decrease in interference. Kanne et al., using the color-word task, found an increase in interference, but from two to three colors, and not from three to four. Since the two-stimuli condition is special, because each color is perfectly correlated with a distractor word, this condition cannot be directly compared to conditions relying on a larger number of colors.

In general, the reasons for unclear results concerning set-size effects in studies administered to date may pertain to the fact that so far no study succeeded in simultaneously addressing all of the following methodological problems: too small sample (i.e., fewer than twenty people per a set-size group), either too small or too large set-size variation, using the two-response condition as a

baseline, using a different material in the small versus large set-size conditions (for an exception see La Heij, van den Hof, 1995; La Heij, Vermeij, 1987), introducing a larger amount of negative priming in the former condition than in the latter condition (a problem also effectively addressed in La Heij's studies), decreasing a number of target-distractor associations in the incongruent trials with an increased set size (see Melara, Algom, 2003), and decreasing a number of trials per target with an increasing target/response set size (accounted for in Kanne et al., 1998). Concluding the review of previous studies, the problem of set-size effects in Stroop has not been resolved yet, and it calls for more data that will enable psychology to establish the factual relation between set size and interference.

However, why should it matter whether the interference effect does or does not increase by only a few tens of milliseconds when participants cope with, let us say, eight stimuli instead of four? The answer is that knowledge of the set-size effect helps to evaluate the predictions of existing theoretical models of interference control within the Stroop task, and knowing which models are plausible and which are not gives insight into the architecture of human cognitive control. Because effective control constitutes a crucial ability for coping with novel and difficult situations, especially when stimulation interferes with internally represented goals, a better understanding of the nature of cognitive control is of primary importance. It is especially highlighted in cases of cognitive control deficits, which yield serious psychological (e.g., perseveration, ineffective planning and coordination of behavior, etc.) as well as social consequences (e.g., violating legal norms, addictions, etc.).

Some existing models of the Stroop effect (e.g., Altmann, Davidson, 2001; Levelt, Roelofs, Meyer, 1999; Roelofs, 2003; van Maanen, van Rijn, Borst, 2009), henceforth referred to as *retrieval competition models*, assume that one and the same group of processes generate the output for color naming and word reading, and the interference effect results 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 are more available, and their retrieval is more probable, but additional activation flowing from the representation of the goal/task boosts the activation of the color-related chunks, and eventually allows for their retrieval, however, yielding an additional processing cost. The key attribute of the retrieval competition models, directly resulting from their architectural assumptions (see Roelofs, 2001; van Maanen et al., 2009), is the fact that the retrieval latency of the proper color-related chunk is 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, and the longer it takes to retrieve the color-related chunk. Because neither neutral nor congruent trials activate distractors, the inherent consequence of the *retrieval competition models* is the prediction that the difference in latency between these two kinds of trials (i.e., the interference effect) increases as the response set size increases (see Roelofs, 2001).

Some other Stroop models (e.g., Cohen, Dunbar, McClelland, 1990; Pfaf, Van Der Heuden, Hudson, 1990; for newer propos-

als see Herd, Banich, O'Reilly, 2006; Smoleń, Chuderski, 2010), henceforth referred to as *response competition models*, assume that interference effects arise from the competition between two structurally separate processing pathways, the dominant one translating the currently displayed word into one possible response, and the weaker one translating the currently displayed color/ picture into the alternative response. Due to the top-down influence of the goal/task, the alternative pathway wins the competition, but at an additional processing cost of overriding the dominant pathway. Computational simulations with the use of some of the *response competition models* showed that the number of possible stimuli or responses does not influence the interference effect (Chuderski, Smoleń, 2011; Cohen, Usher, McClelland, 1998; Kanne et al., 1998). This fact is the inherent consequence of the theoretical assumption of these models, which claim that always only two processing pathways, directly activated by two aspects of a presented stimulus (i.e., the meaning and color of a word), compete for the output, irrespectively of the number of possible stimuli in the task. In other words, *response competition models* locate the Stroop interference at the late stage of processing (at the choice between applicable responses), after the search of and access to task relevant representations (e.g., distractors) have already been carried out. On the other hand, *retrieval competition models* assume that the interference arises at an earlier stage of memory (e.g., lexical) access, when all task relevant representations may potentially influence the amount of arising interference.

The present study aims to investigate whether set-size effects in Stroop actually do or do not exist, and in consequence, depending on the eventual result, it intends to validate one while putting into question the other class of the Stroop models. We plan to compare interference effects in situations differing in the number of stimuli/responses, using both the picture-word (Experiment 1) and color-word (Experiment 2) versions of the Stroop task, while avoiding most of the above mentioned methodological problems of the former studies.

We applied the task, which requires manual responding, because in (most widely used so far) the vocal variants of the Stroop task, interference/conflict resulting from incongruent stimuli was confounded with interference/conflict arising from the simultaneous reading of an incongruent word (and, most probably, silently saying it) and the speaking of a response word (i.e., the effector-specific conflict). Accounting separately for these two sources of interference requires special tasks (see Zhang, Kornblum, 1998), which impose unnatural requirements for the participants (e.g., saying 'green' or 'one' in response to blue color). In the manual variant, interference/conflict are primarily initiated by the conceptual similarity between two aspects of a stimulus: its color and its name (see Kornblum, Lee, 1995), as the conflict between the likely reading/speaking of incongruent word and the pressing of a key is minimal. However, contrary to Kornblum and colleagues (Zhang, Zhang, Kornblum, 1999), for whom this conflict lies in two alternative processing paths leading from stimulus identification to the conceptual level (e.g., to the intermediate layer in their network), we (Smoleń, Chuderski, 2010) and others (Cohen et al., 1990; Roelofs, 2003) define such a conflict as resulting from the competition between paths leading from some in-

termediate level to the response selection stage (e.g., depending on the model cited, being the process of selecting response rule, output node, or lemma, respectively). Thus, there are good reasons to assume that the manual variant of the Stroop tasks allows for the investigation of the crucial, decisional stage of interfering information processing, while it escapes the effector-specific conflicts (probably less relevant for research on cognitive control).

Furthermore, all previous studies, which manipulated response set sizes (especially those using the vocal task) consequently used one-to-one mappings between stimuli and responses, so they were not able to deconfound the perfectly correlated numbers of stimuli and responses. Here, we also examined a novel four-response condition with two-to-one stimuli-response mappings (Experiment 1) in order to separately investigate the response set-size and stimuli set-size effects.

EXPERIMENT 1

Participants

Volunteer participants were recruited via publicly accessible social networking websites. Each participant gave an informed consent and was paid the equivalent of 6 euros in Polish zloty. A total of 177 women and 105 men participated. The mean age was 22.7 years (SD = 4.37, range 16 – 44). All participants had normal or corrected-to-normal vision. Before analyzing the data, 3 participants were excluded for failing the 60% mean accuracy criterion. Participants were randomly assigned to three groups (from 61 to 64 each); the fourth group of 92 people was investigated afterwards.

Materials and Procedure

We examined Stroop conditions, which included four (4S-4R), six (6S-6R), or eight (8S-8R) stimuli and corresponding responses. In each condition, we presented a random sequence including 20 congruent and 20 incongruent trials per stimulus, which was always a geometric figure including a word. This led to 160 trials in the 4S-4R condition, 240 trials in the 6S-6R condition, and 320 trials in the 8S-8R condition. We used a constant number of trials, instead of a constant sequence length, because the level of automatization in a task probably depends on the number of encounters with each particular stimulus (and distractor) and not on the total number of performed trials (Logan, 1988). In the congruent trials, a figure included a word naming this very figure, while in the incongruent trials figures included words naming different figures. The task was to name a figure while ignoring the associated word. Importantly, the same pool of eight figures (listed below) was used in each condition. In the 4S-4R and 6S-6R conditions, each participant saw a random subset of this pool. In each condition, 80 training trials were applied. Primarily, we aimed to compare the interference effects between the three conditions of the task.

In the 8S-8R trials, eight shapes were used: a rectangle, rhombus, cross, trapezium, splash, oval, circle, and ring. The first four shapes fell in the category of "angular shapes", while the last four formed the "round shapes" category (see below). Each shape was approximately 6 × 5 cm in size, and was presented in blue with a black outline on a dark grey background. Each stimulus included a capital word in the middle of

it, either congruent (e.g., the circle shape including word "CIRCLE", in Polish) or incongruent (e.g., the circle shape including word "CROSS", in Polish), approximately 3.5×0.8 cm in size, printed in black. Each stimulus was randomly associated with one of the following key-finger mappings: "q" – the left little, "w" – left ring, "e" – left middle, "r" – left index, "u" – right index, "i" – right middle, "o" – right ring, and "p" – right little finger. In the 6S-6R trials, six shapes were randomly selected and randomly associated with "q" – left ring, "w" – left middle, "e" – left index, "i" – right index, "o" – right middle, and "p" – right ring fingers. In the 4S-4R trials, four random shapes were picked and randomly associated with "q" – left middle, "w" – left index, "o" – right index, and "p" – right middle fingers.

Finally, we tested an eight-stimuli condition including two-to-one S-R mappings (8S-4R). In these trials, participants used the index and middle fingers of each hand mapped to the keys exactly as in the 4S-4R trials, but two shapes were mapped to each finger. Half the participants encountered random stimuli-finger associations, while in the other half, two random angular shapes were associated with the left middle finger, two other angular shapes – with the left index finger, two random round shapes – with the right index finger, and two other round shapes – with the right middle finger.

In each trial, firstly, a fixation symbol was presented for 2 s in the center of the computer screen, then a random shape from the stimuli set (with the constraint that the same shape cannot be directly repeated) was presented in this very place until either a response was given or maximum of 2.5 s passed, and finally a mask was shown for 1 s. The mask indicated a pause before the next trial occurred. Detailed instructions preceded the training and the test.

Hypotheses and Data Analysis

The latency of correct responses was our primary dependent variable. We screened for latencies shorter than 250 ms (0.65% of data). Latencies longer than 2500 ms were prevented by the experimental procedure. Most importantly, for each participant we calculated the interference effect as the difference between the mean RT in the incongruent trials and the congruent trials.

We tested two alternative hypotheses predicted by either the *retrieval or response competition models.* The former models predict a significant difference in the interference effect between at least two set-size conditions, with a larger set size yielding larger interference, in comparison to a smaller set size. In contrast, the response competition models predict that no significant difference in interference can be found between any set-size conditions. Moreover, if the former hypothesis appears supported, for example, the 8S-8R condition results in a significantly larger interference effect than the 4S-4R condition, then testing the 8S-4R condition may tell us whether either the increase in the number of stimuli or involving more responses (or both) is responsible for the observed increase in interference. The former conclusion can be made if the 8S-4R condition yields a significantly larger interference effect than the 4S-4R condition, meaning that increasing solely the number of stimuli (from four to eight), with constant response set size (equaling four), leads to an increased interference. The latter inference will be justified if the 8S-4R condition yields a significantly smaller interference effect than the 8S-8R condition,

as only the number of responses increases (from four to eight) between these conditions, while the stimuli set size (equaling eight) is kept constant. The last possible outcome regarding set-size effects pertains to a decreased interference when the set size increases, however, no model known to us predicts such a result.

We statistically controlled for a number of possible confounding factors, but it must be noted that it is impossible to account for all such factors simultaneously. For example, although the figures in the adopted experimental design were perfectly uncorrelated with the words in the incongruent trials, the respective correlation in all trials (congruent and incongruent) differed between conditions (see Melara, Algom, 2003), and equaled $r_{4R} = .33$, $r_{6R} = .40$, and $r_{8R} = .43$. In order to balance these correlations, the proportion of congruent trials should have been decreased in the 6S-6R and 8S-8R conditions, but such a manipulation would also decrease the interference effects (Logan, Zbrodoff, 1979). In particular, in the 8S-8R condition, achieving zero correlation should have required including only 12.5% of the congruent trials, but that would have led to drastically reduced interference effects. Thus, using the same proportions of congruent and incongruent stimuli (in our study: 50/50) in all compared conditions seems to be the best methodological choice.

Nevertheless, we de-confounded a number of other factors, which were mentioned above. Firstly, in subsequent analyses, we compared 160 trials from the 4S-4R condition to the first 160 trials from each of the two remaining conditions. Secondly, we included in the analyzed data the last 40 and 20 training trials from the 4S-4R and 6S-6R condition respectively, but excluded the last 40 and 20 test trials respectively, thus balancing the number of training trials to 10 per stimulus. Thirdly, as the amount of negative priming (e.g., the probability of the n-1-trial distractor directly preceding the figure denoted by this very distractor) decreased as the response set size increased, we excluded from analysis all negative-priming trials. Fourthly, as the number of particular figure-word associations in incongruent trials decreased from the 4S-4R (20/3), through the 6S-6R (20/5), to the 8S-8R condition (20/7), we randomly picked only two trials for each such association, we picked up three associations for each of four figures (in the 6S-6R and 8S-8R conditions, we picked random associations and figures), and calculated the interference effects from the resulting 24 incongruent and 24 congruent trials, thus accounting for the equal number of figure-word associations.

The additional manipulation in shape-finger associations (i.e., arbitrary vs. univocal mapping) in the 8S-4R condition was aimed to test whether the univocal mapping (i.e., grouping stimuli into meaningful categories) decreases the interference effect in comparison to the arbitrary (non-meaningful) mapping. This could indicate that working memory affects interference, as in the univocal condition, the working memory load should be lower than in the arbitrary condition, probably due to a better chunking in working memory of S-R mappings with regard to each hand, or because of an effective precuing of the selection of S-R mappings to working memory by a salient feature (i.e., either angularity or roundness) of a stimulus (see Adam, Hommel, Umiltá, 2003; Proctor, Reeve, 1985).

Having a relatively large sample of participants, we also tested for any gender differ-

ences in interference effects, since sometimes (though not always; see MacLeod, 1991) it has been found that women perform better than men (e.g., Sarmány, 1977). Maybe, a difference in interference shows up when relatively many (i.e., eight) responses are needed.

RESULTS

Because we aimed at showing the null hypotheses to be true, while in frequentist statistics, in principle, it cannot be decisively shown that they cannot be rejected (Rouder et al., 2009). In the following analyses we used the Bayes factor (K statistics, Kass, Raftery, 1995), which is the ratio between posterior probabilities of the considered models. In the case of the results reported below, K is the ratio of the posterior probability of the model, assuming an effect (M_1) to the posterior probability of the null model (M_0). In general, the K value is the extent to which the ratio of the two, initially equal, prior probabilities for the two compared models should be updated on the basis of the observed data, in order to obtain the appropriate ratio of posterior probabilities. According to Kass and Raftery, if K < 1, then M_1 should be decisively rejected, and M_0 – accepted; if K < 3 then there is little evidence for M_1; and when K > 3, then M_1 can be accepted (e.g., K = 3 means that we should assign M_1 a posterior probability three times larger than that of M_0). The "BayesFactor" package in R (Morey, Rouder, 2013) was used.

Below, we also applied multiple F tests, but taking into consideration the fact that our hypotheses pertained to null effects, we did not run any multiple comparison correction. Such a procedure constitutes a more conservative strategy than applying such corrections, because it minimizes the probability of type II error, thus, in the present case it seems fully justified.

Comparison of the 4S-4R, 6S-6R, and 8S-8R conditions. The mean error rate was very low (M = .034), and thus was not further analyzed. The mean latency increased significantly with the increasing number of possible responses, in cases of both congruent, $F(2, 184) = 24.52$, $p < .001$, $\eta^2 = .20$, and incongruent trials, $F(2, 184) = 22.38$, $p < .001$, $\eta^2 = .19$ (see Table 1), reflecting the well-known Hick's law (Hick, 1952). However, no significant difference between conditions was found in the case of the interference effect (see Figure 1), $F(2, 184) = 0.89$, K = 0.12. The power of our experiment to reject the null effect was $1 - \beta = .67$, which means that if the real null effect was not found in the data, then it would have been rejected with a .67 chance, but it was not. Adding the gender factor to the model did not improve its fit, $F(3) = 0.98$, K = 0.06.

In subsequent analyses, we controlled for the above mentioned, potentially confounding variables. A direct comparison of the interference effects in the first 160 trials in each set-size condition did not yield any significant difference, $F(2, 184) = 0.89$, K = 0.12.

Table 1. Means and standard deviations of reaction times (in millisec.) in Experiment 1

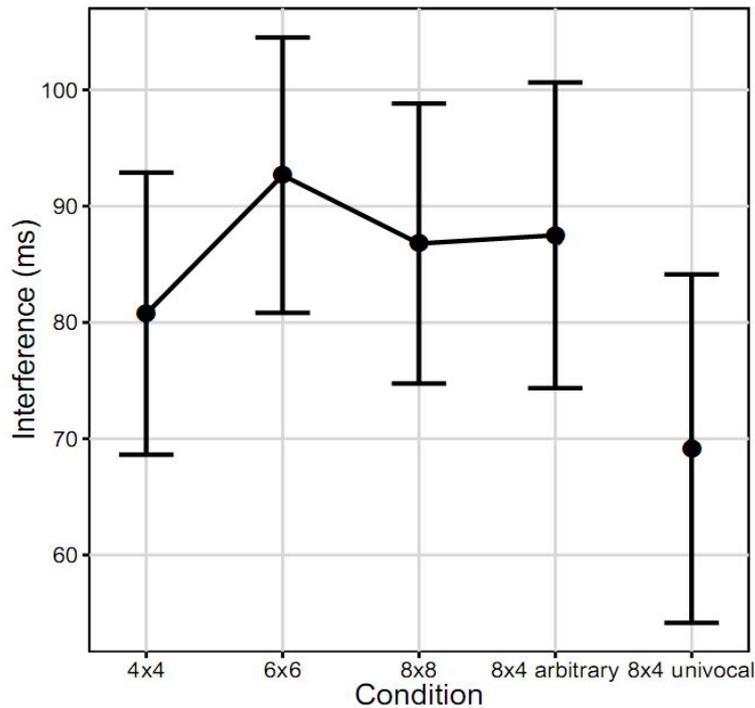|	Task condition			
	4S-4R	6S-6R	8S-8R	8S-4R
Congruent trials	784 (161)	852 (157)	979 (152)	871 (172)
Incongruent trials	865 (167)	945 (173)	1065 (162)	951 (193)

Figure. 1. The comparison of the latency interference effects (i.e., the differences in the mean latency of correct responses between incongruent and congruent trials) in Experiment 1, between the four-stimuli/four-responses (4S-4R), six-stimuli/six-responses (6S-6R), eight-stimuli/eight-responses (8S-8R), and arbitrary-mapping eight-stimuli/four-responses (8S-4R) variants of the Stroop task. None of the respective differences was significant at the $p < .05$ level. The univocal-mapping 8S-4R condition, which possibly involved the reduced WM load, and as the only one yielded significant differences, is presented as an isolated point at the rightmost edge. Bars indicate 95% of confidence intervals.

Neither did equating the training sequence length (to 40 trials) in the three conditions, and testing the subsequent 160 trials, $F(2, 184) = 0.77$, $K = 0.10$, nor testing the subsequent 160 trials (in 4S-4R condition) vs. 240 (6S-6R) vs. 320 (8S-8R) ones, $F(2, 184) = 0.82$, $K = 0.11$. The differences between interference effects were also non significant, $F(2, 184) = 1.03$, $K = 0.13$, when the negative-priming trials were excluded. Finally, randomly picking only two trials for each figure-word association did not yield significant set-size effects either, $F(2, 184) = 0.08$, $K = 0.06$.

Comparison of the 8S-4R and 8S-8R conditions (testing the sheer response set-size effect). The error rate in the 8S-4R condition ($M = .066$) was at the bottom. Again, the mean latency increased significantly with the increasing number of possible responses, in the case of both congruent, $F(1, 152) = 15.84$, $p < .001$, $\eta^2 = .09$, and incongruent trials, $F(1, 152) = 14.86$, $p < .001$, $\eta^2 = .08$ (see Table 1). However, again, there was no significant response set-size effect regarding the interference effect (see Figure 1), $F(1, 152) = 1.00$, $K = 0.28$. Again, adding the gender factor did not improve the fit of the model, $F(2) = 0.41$, $K = 0.07$. A direct comparison of interference effects between the arbitrary and univocal mapping conditions within the 8S-4R group revealed a marginal effect, $t(90) = 1.96$, $p = .052$, $K = 1.19$, indicating that applying more meaningful S-R mappings helped somewhat to decrease interference ($\Delta = 18$ ms). Only the univocal mapping condition yielded a significant difference in the interference effect (69 vs. 87 ms), compared to the 8S-8R condition, $t(100) = 2.07$, $p = .041$, while the difference between the arbitrary 8S-4R condition and the 8S-8R condition (88 vs. 87 ms) was non significant, $t(112) = 0.08$, $K = 0.20$, no matter if negative priming was accounted for or not.

Comparison of the 4S-4R and 8S-4R conditions (testing the sheer stimuli set-size effect). Finally, we compared the interference effects in the 4S-4R and 8S-4R conditions (see Figure 1). There was barely a difference in the latency interference effect (Ms = 81 and 80 ms, respectively), and it was non significant in both the arbitrary mapping 8S-4R condition (81 vs. 88 ms), $t(111) = 0.70$, $K = 0.25$, and the univocal mapping 8S-4R condition (81 vs. 69 ms), $t(99) = 1.18$, $K = 0.40$. The difference was negligible even if we used only the first 160 trials of the 8S-4R condition, $t(151) = 0.15$, $K = 0.18$, equated the number of training trials, $t(151) = 0.19$, $K = 0.18$, and excluded negative-priming trials, $t(151) = 0.26$, $K = 0.18$.

DISCUSSION

Our results, at least with regard to four versus eight elements, indicate that Stroop interference does not depend on the set size. They also shed some new light on the problem of contradictory data found in the existing studies (see La Heij, van den Hof, 1995; MacLeod, 1991; Spieler et al., 1996). Moreover, our study de-confounded in a new way the response set-size effects from the stimuli set-size effects, and revealed that there were neither the former nor the latter set-size effects. Because we used large samples in each experimental condition, and we controlled for most of potentially confounding variables, our findings seem to be highly reliable.

Only the univocal mapping 8S-4R condition yielded a marginally significant decrease in the interference effect. However, this is probably only evidence for the fact that the interference effect is partially related to the working memory load, and that chunking or precuing S-R mappings due to meaningful categories helps to decrease such a load. Nevertheless, the decrease in the load affected interference to relatively limited extent, so it seems that the need to maintain in memory the S-R mappings in the Stroop task is not a primary cause for interference effects to appear in this task.

However, one objection to the present study may be that the picture-word task is known to yield relatively lower interference effects than does the classic color-word task. So, in the former version, the conflicts

between representations of responses occurring at a retrieval stage may not be strong enough to yield a difference in interference when the set of such representations increases. The next experiment was aimed to test if such a difference can appear when stronger Stroop interference is induced by the color-word task. Expecting to replicate null results found in Experiment 1, we tested only the extreme, 4S-4R vs. 8S-8R conditions.

EXPERIMENT 2

Participants

The same recruitment and testing conditions and gratification were applied as in Experiment 1. A total of 79 women and 44 men participated. The mean age was 24.41 years (SD = 5.15, range 19 – 45). Before analyzing the data, 4 participants were excluded for failing the 60% mean accuracy criterion. Participants were randomly assigned to the 4S-4R (60 people) or 8S-8R (59 people) groups.

Materials and Procedure

The Stroop conditions, which included four (4S-4R) or eight (8S-8R) stimuli and corresponding responses, were used exactly as in Experiment 1, with two exceptions: each group received 40 training trials (as their exact number had a negligible effect in Experiment 1), and the stimuli were words in capital letters (approx. 8×3 cm in size), meaning eight colors (black, red, green, blue, yellow, pink, brown, orange), displayed in either congruent colors (e.g., word "RED", in Polish, printed in red color) or incongruent colors (e.g., word "RED", in Polish, printed in blue color).

RESULTS AND DISCUSSION

In Experiment 2, the same statistical procedure regarding K values was adopted as in Experiment 1. The error rate was again at the bottom (M = .028). The mean latency increased significantly when the number of possible responses increased from four to eight, in the case of both congruent, $F(1, 117) = 76.9$, $p < .001$, $\eta^2 = .39$, and incongruent trials (see Table 2), $F(1, 117) = 67$, $p < .001$, $\eta^2 = .36$, but there was virtually no numerical difference between the interference effects in the 4S-4R (M = 102 ms) and 8S-8R conditions (M = 105 ms), $F(1, 117) = 0.08$, $p = .77$, K = 0.20 (see Figure 2). The observed power of the test equaled $1 - \beta = .78$. Adding the gender factor did not improve the fit of the model, $F(2) = 1.75$, K = 0.25. A direct comparison of interference effects did not yield a significant effect even if we used only the first 160 trials of the 8S-8R condition, $t(117) = 0.39$, K = 0.21, or excluded negative-priming trials, $t(117) = 0.86$, K = 0.28. Thus, the lack of set-size effect in the manual Stroop task seems to have been reliably replicated.

Table 2. Means and standard deviations of reaction times (in millisec.) in Experiment 2

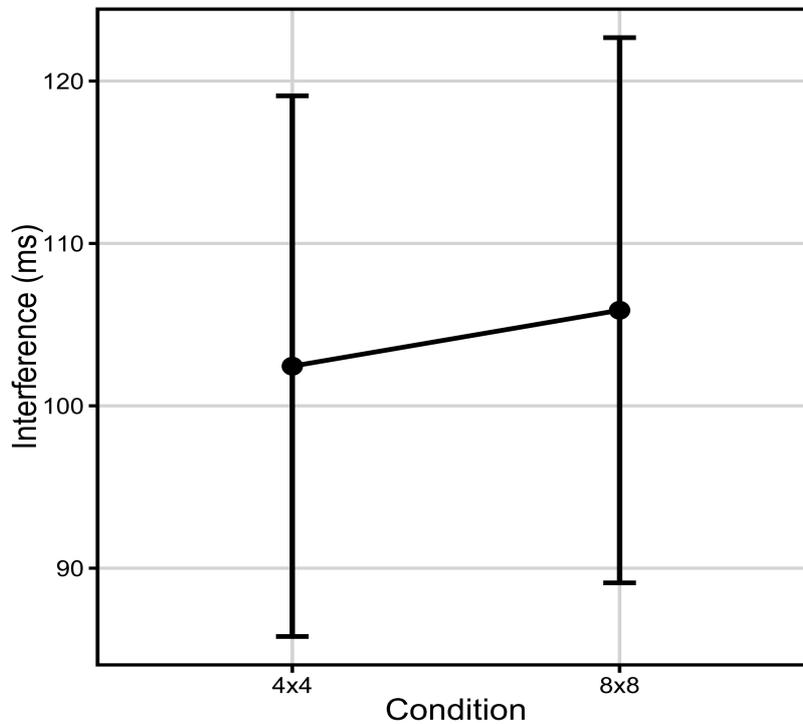|	Task condition	
	4S-4R	8S-8R
Congruent trials	819 (188)	1095 (154)
Incongruent trials	921 (205)	1202 (166)

Figure. 2. The comparison of the latency interference effects (i.e., the differences in the mean latency of correct responses between incongruent and congruent trials) in Experiment 2, between the four-stimuli/four-responses (4S-4R) and eight-stimuli/eight-responses (8S-8R) variants of the Stroop task. This difference was not significant at the p < .05 level. Bars indicate 95% of confidence intervals.

GENERAL DISCUSSION

We showed that neither stimuli nor response set-size effects were found in the two most popular versions of the manual Stroop task. As both our experiments had sufficient power to detect set-size effects if they really existed, our study seems to help resolve the contradiction among existing research, probably rooted in the fact that not all studies succeeded in de-confounding set-size manipulations from other factors, for instance from the working memory load, which, as we showed by means of the univocal condition of Experiment 1, can have a certain influence on interference.

The manual Stroop is obviously limited by the number of possible responses, and the same applies to our conclusion, as some vocal versions of the task, including more responses, have been shown to yield sub-

stantial response set-size effects (e.g., La Heij, van den Hof, 1995). It is possible that set-size effects would appear if our set size was increased to 16 or more S-R mappings, compared to the 8 mappings used in our study. On the other hand, in cases of such large alteration of testing conditions, it may not be possible to attribute set-size effects solely to the manipulation applied. As discussed above, fulfilling the more difficult multiple-stimuli/response task may involve a larger load on working memory capacity. It may also increase task related arousal/motivation, or yield plenty of other differences in comparison to the simpler, few-stimuli/response versions. If the set-size effect were a real phenomenon, it should have shown up even under conditions when the set size was doubled, but no such result has been found in our data.

One reason for set-size effects sometimes appearing in the vocal Stroop but absent in the present manual Stroop, may be that the former yields high similarity between stimuli and response sets (dimensional overlap between them, see Kornblum, Lee, 1995), while the latter, involving key presses to figures or colors, yields relatively dissimilar mappings (i.e, no overlap). Thus, potential vocal responses may resemble distractors to such a large degree that they lead to partial activation of the entire response set. Key presses, which resemble neither targets nor distractors, may require additional translation from stimulus to response and thus the entire response set may not be activated, rather, only individual competing responses may be affected. So, vocal and manual versions of the Stroop task may involve to some extent different cognitive mechanisms. However, vocal tasks, which may not require such a translation, seem to confound, as noted above, interference resulting from incongruent stimuli with the effector-specific conflict. Thus, manual tasks may be a tool better suited for the verification of predictions of both *retrieval and response competition models*, which pertain primarily to conflicts at the decisional level, and not at the response-generation level (in fact, they barely account for the latter conflicts). However, our next step in research on set-size effect in Stroop may include its careful examination by using vocal tasks.

Finally, our pattern of results is inconsistent with the central prediction of the *retrieval competition models*, which explain interference in terms of semantic conflict between all alternative S-R mappings. Although it is possible that the semantic mechanisms related to word retrieval can affect the basic conflict resolution mechanisms, and result in, for example, semantic interference (La Heij, van den Hof, 1995) or semantic gradient effects (MacLeod, 1991), it seems that (contrary to La Heij and van den Hof's observation) the lion's share of the interference effect, at least in the manual versions of the Stroop task, is due to competition between the established processing paths. In contrast, our results are in concord with the inherent predictions of the *response competition models* (Chuderski, Smoleń, 2011; Cohen, Usher, McClelland, 1998; Kanne et al., 1998), and suggest that conflicts between cognitive operations resulting in interference may be elicited primarily by stimuli, which are currently perceived and which directly activate alternative response tendencies, and not by the remaining targets/distractors/response options. Although one type of test of the models' predictions can never lead to the definite acceptance of some and the rejection of other models, and more future work

is needed for the comprehensive evaluation of the theoretical models of cognitive control phenomena, the present study seems to have helped to identify which models of Stroop are more or less successful.

REFERENCES

ADAM, J.J., HOMMEL, B., UMILTÁ, C., 2003, Preparing for perception and action (I): The role of grouping in the response-cueing paradigm. *Cognitive Psychology,* 46, 302-358.

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.

CHUDERSKI, A., SMOLEŃ, T., 2011, The effect of response set size on the Stroop interference. In: C. Hoelscher, T.F. Shipley, L. Carlson (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2162-2167). Austin, TX: Cognitive Science Society.

COHEN, J.D., DUNBAR, K., MCCLELLAND, J.L., 1990, On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review,* 97, 332-361.

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.

DANIEL, J., 1970, Further variants of Stroop's interference test. *Studia Psychologica,* 12, 80-81.

GLASER, M.O., GLASER, W.R., 1982, Time course analysis of Stroop phenomenon. *Journal of Experimental Psychology: Human Perception and Performance*, 8, 875-894.

HERD, S.A., BANICH, M.T., O'REILLY, R.C., 2006, Neural mechanisms of cognitive control: An integrative model of Stroop performance and fMRI data. *Journal of Cognitive Neuroscience,* 18, 22-32.

HICK, W.E., 1952, On the rate of gain of information. *Quarterly Journal of Experimental Psychology,* 4, 11-26.

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.

KASS, R.E., RAFTERY, A.E., 1995, Bayes factors. *Journal of American Statistics Association*, 90, 773-795.

KORNBLUM, S., LEE, J.W., 1995, Stimulus-response compatibility with relevant and irrelevant stimulus dimensions that do and do not overlap with the response. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 855-875.

LA HEIJ, W., VAN DER HEIJDEN, A.H.C., SCHREUDER, R., 1985, Semantic priming and Stroop-like interference in word-naming tasks. *Journal of Experimental Psychology: Human Perception and Performance,* 11, 62-80.

LA HEIJ, W., VAN DEN HOF, E., 1995, Picture-word interference increases with target-set size. *Psychological Research,* 58, 199-133.

LA HEIJ, W., VERMEIJ, M., 1987, Reading versus naming: The effect of target set size on contextual interference and facilitation. *Perception & Psychophysics,* 41, 355-366.

LEVELT, W.J.M., ROELOFS, A., MEYER, A.S., 1999, A theory of lexical access in speech production. *Behavioral and Brain Sciences,* 22, 1-38.

LOGAN, G.D., 1988, Toward an instance theory of automatization. *Psychological Review,* 95, 492-527.

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.

MACLEOD, C.M., 1991, Half a century of a research on the Stroop Effects: An integrative review. *Psychological Bulletin,* 109, 163-203.

MACLEOD, C.M., DUNBAR, K., 1988, Training and Stroop-like interference: Evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 15, 126-135.

MELARA, R.D., ALGOM, D., 2001, Driven by information: A tectonic theory of Stroop effects. *Psychological Review,* 110, 422-471.

MOREY, R.D., ROUDER, J.N., 2013, *BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.5.* [Computer software manual] Retrieved from: http://CRAN.R-project.org/package=BayesFactor

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.

PROCTOR, R.W., REEVE, T.G., 1985, Compatibility effects in the assignment of symbolic stimuli to discrete finger responses. *Journal of Experimental Psychology: Human Perception and Performance,* 11, 623-639.

RAY, C., 1974, The manipulation of color response times in a color-word interference task. *Perception and Psychophysics,* 16, 101-104.

ROELOFS, A., 2001, Set size and repetition matter: Comment on Caramazza and Costa (2000). *Cognition,* 80, 283-290.

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

ROUDER, J.N., SPECKMAN, P.L., SUN, D., MOREY, R.D., IVERSON, G., 2009, Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review,* 16, 225-237.

SARMÁNY, I., 1977, Different performance in Stroop's interference test from the aspect of personality and sex. *Studia Psychologica,* 19, 60-67.

SMOLEŃ, T., CHUDERSKI, A., 2010, Modeling strategies in Stroop with a general architecture of executive control. In: S. Ohlsson, R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 931-936). Austin, TX: Cognitive Science Society.

WILLIAMS, E., 1977, The effects of amount of information in the Stroop color word test. *Perception and Psychophysics,* 22, 463-470.

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.

ZHANG, H., KORNBLUM, S., 1998, The effects of SR mapping, and irrelevant SR and SS overlap in four-choice Stroop tasks with single carrier stimuli. *Journal of Experimental Psychology: Human Perception and Performance,* 24, 3-19.

ZHANG, H., ZHANG, J., KORNBLUM, S., 1999, A parallel distributed processing model of stimulus-stimulus and stimulus-response compatibility. *Cognitive Psychology,* 38, 386-432.

POČET ODPOVEDÍ/STIMULOV NEMÁ VPLYV NA MANUÁLNY STROOPOV TEST

A. C h u d e r s k i, T. S m o l e n, M. T a r a d a y

Súhrn: Existujúce výpočtové modely Stroopovho testu sa líšia v predpovediach týkajúcich sa vplyvu veľkosti vzorky, t.j. vzťahu medzi počtom stimulov/odpovedí a veľkosťou Stroopovej interferencie. Relevantné empirické dáta však nie sú jednoznačné, keďže niektoré štúdie nezaznamenali vplyv veľkosti vzorky, zatiaľ čo iné vykazovali podstatný vplyv veľkosti vzorky. Aby sme vyriešili túto diskrepanciu, urobili sme dva experimenty s manuálnym Stroopovým testom. Experiment 1 porovnával podmienky so štyrmi, šestimi a ôsmymi stimulmi/odpoveďami v obrázkovo-slovnej úlohe. Nezistili sme žiadne relevantné vplyvy veľkosti vzorky, okrem malého efektu, keď sa vplyvom úlohy zámerne znížilo zaťaženie pracovnej pamäti. Experiment 2 testoval podmienky štyroch vs. ôsmich mapovaní vo farebno-slovnej úlohe a replikoval výsledky Experimentu 1. Keďže oba experimenty by dokázali odhaliť vplyv veľkosti vzorky pokiaľ by tam bol, naše dáta sa nezhodujú s modelmi, ktoré predikujú takýto efekt.