The Classic Stroop Asymmetry in Online Experiments
Mary E. Smith, Micah D. Smith, and Kenith V. Sobel*
Department of Psychology and Counseling, University of Central Arkansas

ABSTRACT. In a traditional Stroop (1935) task, participants view target words written in colors that are either congruent with their meaning (e.g., “Red” written in red ink) or incongruent (e.g., “Red” written in green). When participants identify the target color, the response time difference between congruent and incongruent targets (i.e., Stroop effect) is typically much larger than when participants identify the target word (i.e., reverse Stroop effect); this is the classic Stroop asymmetry. Recent work has shown that recasting the task so participants localize the target rather than identify it inverts the asymmetry: the Stroop effect is smaller than the reverse Stroop effect. We developed online identification and localization scripts in an attempt to replicate the classic Stroop asymmetry and this recent twist on the classic asymmetry. In Experiment 1, a sample of undergraduate students replicated the classic Stroop asymmetry with an identification task and its inversion with a localization task, \( p < .001, \eta_p^2 = .17 \). In Experiment 2, a sample of participants who were more representative of the general population than in Experiment 1 once again replicated the classic Stroop asymmetry with an identification task and its inversion with a localization task, \( p < .001, \eta_p^2 = .20 \). Our results combine with other recent studies to provide converging evidence that the classic Stroop asymmetry results from the strength of association between task demands (identification vs. localization) and the attended feature (verbal vs. visual).

Keywords: Stroop effect, reverse Stroop effect, Stroop asymmetry, visual search, online data collection
Here we describe experiments based on Qualtrics scripts that attempt to replicate both the classic Stroop asymmetry and a recent study that inverted the classic asymmetry. The venerable Stroop (1935) paradigm is ideal for validating a novel method of capturing RT (Barnhoorn et al., 2015; Grump et al., 2013), because Stroop effects have been so reliably replicated (MacLeod, 2005).

The Classic Stroop Asymmetry
In a typical computer-based Stroop task, participants view a color word written in pixels that either match (e.g., “Red” in red pixels) or do not match (e.g., “Red” in green pixels) its meaning. Participants either report the target color while ignoring its meaning (Stroop condition), or the target word while ignoring its color (reverse Stroop condition). When the unattended feature is incongruent with the attended feature, it can leak through the attentional filter to interfere with participants’ ability to report the attended feature (Linzarini et al., 2017). The difference in RT between congruent and incongruent targets represents the magnitude of the interference (Whitehead et al., 2018). Interference from the unattended feature is typically greater in the Stroop condition, in which participants report the target color, than in the reverse Stroop condition, in which they report the target word; this is the classic Stroop asymmetry (Melara & Algom, 2003).

One prominent explanation for the classic Stroop asymmetry asserts that visual and verbal information are encoded and processed in two different systems (Song & Hakoda, 2015; Virzi & Egeth, 1985), and interference can occur when the stimulus is encoded in one format while the response is encoded in the other. For example, in the traditional Stroop condition, participants vocally report the target color, so the visually encoded stimulus (target color) must be translated into a verbal code in order to vocalize it (Durgin, 2000). Even if the response is manual rather than vocal, as when participants press a key to report the target color, the manual response is verbally mediated (Bearden et al., 2021; Blais & Besner, 2006; Parris et al., 2019; Sugg & McDonald, 1994). Whereas translation is required when participants report the target color in the Stroop condition, when participants report the target word in the reverse Stroop condition, no translation is required between the (verbally encoded) target word stimulus and the (verbally mediated) response. Thus, the translation account explains why the Stroop effect is typically larger than the reverse Stroop effect, but it also predicts that if the response were changed to require a visual rather than verbal code, translation would be required in the reverse Stroop condition between the (verbally encoded) target word stimulus and the (visually encoded) response. As a result, the asymmetry would be inverted: the Stroop effect should be smaller than the reverse Stroop effect.

Durgin (2000; replicated by Miller et al., 2016) tested this prediction of the translation hypothesis. Each display contained a target word surrounded by four rectangular patches of color. The target word was either a color word with an incongruent pixel color, or neutral (noncolor words with colored pixels in the Stroop condition or color words written with gray pixels in the reverse Stroop condition). Participants moved the cursor from the target toward the color patch that was consistent with either the target color in the Stroop condition or the target word in the reverse Stroop condition. Translation was not required in the Stroop condition between the (visually encoded) target color stimulus and the (visually encoded) color patch response, but translation was required in the reverse Stroop condition between the (verbally encoded) target word stimulus and the (visually encoded) color patch response. Consistent with the translation account, the Stroop effect was smaller than the reverse Stroop effect: an inversion of the classic Stroop asymmetry.

Although the results from Durgin’s (2000) color matching experiment were consistent with the translation account, Blais and Besner (2007) identified a potential confound in Durgin’s method. While it is true that Durgin encouraged visual processing in a Stroop task by using visually encoded color patch responses, the task he designed also relied on visual processing. Whereas the traditional Stroop paradigm requires participants to identify the target color or word, Durgin required participants to localize the color patch matching the target color or word. Blais and Besner proposed that traditional identification tasks are more strongly associated with verbal than visual processing, but Durgin’s localization task is more strongly associated with visual than verbal processing. Thus, the strength-of-association account implies that an identification task should confer an advantage on the target word, whereas a localization task should confer an advantage on the target color, so manipulating the task may be sufficient to invert the Stroop asymmetry even without any need for translation.
Sobel et al. (2020) developed a variation on Durgin (2000) and manipulated the task to verify this prediction of the strength-of-association account. Participants identified the target in the first experiment and localized it in the second. As expected, the identification task in the first experiment replicated the classic Stroop asymmetry. Then for the localization task in the second experiment, the center of each display was occupied by a cue indicating what participants should search for on that trial: a rectangular color patch in the Stroop condition, or a color word written in a neutral color in the reverse Stroop condition. Surrounding the cue were four color words that were all either congruent or incongruent with their pixel color. Participants reported the location of the item that had the same pixel color as the color patch cue in the Stroop condition or the same meaning as the word cue in the reverse Stroop condition. Because the cued feature and the relevant target feature relied on the same type of code (i.e., color patch cue and target color in Stroop, word cue and target word in reverse Stroop), there was no need for translation. Nevertheless, as predicted by the strength-of-association account, the localization task itself was sufficient to invert the classic Stroop asymmetry: The Stroop effect was smaller than the reverse Stroop effect.

Hypotheses
For this project, we aimed to develop a Qualtrics script that replicates the classic Stroop asymmetry for an identification task, primarily to demonstrate that the script is sufficiently powerful to detect the effect even with the uncertainties inherent in a remotely delivered task. Also, we aimed to extend on this validity check by attempting to replicate the inversion of the classic Stroop asymmetry with a localization task as in Sobel et al. (2020). Experimental psychologists commonly rely on samples comprising undergraduate students because this population is readily accessible, but the results from studying such convenience samples may not generalize to the overall population (Hanel & Vione, 2016; Peterson & Merunka, 2014). For that reason, after releasing the script to a convenience sample of undergraduate students, we wanted to release it to a sample of participants who are more representative of the general population.

We had two hypotheses. First, we hypothesized that, for a convenience sample of university students in Experiment 1, the identification task would replicate the classic Stroop asymmetry, and the localization task would replicate Sobel et al. (2020) by inverting the asymmetry. And second, we hypothesized that the results from Experiment 1 would generalize to a more representative sample in Experiment 2.

Experiment 1a: Identification Task With University Students

Method
Participants
We submitted an application for review entitled Color and Meaning that was approved by our university’s Institutional Review Board, and treated all participants in accordance with the ethical standards established by the American Psychological Association (2017). In Experiments 1a and 1b, all participants were undergraduate students at a mid-sized school in the southern United States and received credit in a variety of psychology courses in exchange for their participation. We used the results from a pilot experiment to determine an appropriate sample size that would reliably detect a difference between a Stroop effect and reverse Stroop effect. The pilot experiment yielded a Cohen’s $d$ of 0.21, which would require a sample size of 79 participants to achieve 80% power at an alpha of .05 (Bausell & Li, 2002). Our sample for Experiment 1a included 83 participants, 13 of whom identified as male, 69 as female, and one as nonbinary. Participants’ ages ranged from 18 to 61 with a mean of 22.07 and standard deviation of 6.85. There were 62 participants who described themselves as White, 15 as Black or African American, three as Hispanic, one as Asian, one as American Indian or Alaska Native, and one preferred not to report a race. The Qualtrics script randomly assigned each participant to carry out either the identification task (Experiment 1a) or the localization task (Experiment 1b). All participants were exposed to a Stroop condition in one block and a reverse Stroop condition in the other block. Block order was counterbalanced across participants.

Apparatus
Participants logged in to their accounts on Sona (https://UCA.sona-systems.com) to sign up for the experiment. The Sona website then redirected them to the script on Qualtrics, which they could complete by using their own personal devices.

Stimuli
Qualtrics scripts typically require participants to answer one or more questions on each page, then move on to the next page by clicking the Advance
button located in the lower right corner of each page. However, in our study, if the participant’s cursor (for anyone using a personal computer) or finger (for anyone using a tablet device) were situated in the lower right section of the display at the beginning of each trial, we thought they would respond more quickly to response buttons near the Advance button on the right side of the display. To mitigate these biasing effects, we wanted to force participants to position their cursor or finger in the middle of the display just before the response buttons appeared. To do so, each trial began with a page containing a prompt (i.e., “Click here to start the next trial”) above a button situated in the middle of the display. If participants tried to click the Advance button on the lower right side of the display rather than the cursor-centering button, the Qualtrics script insisted that they select the cursor-centering button before they could advance. To make the script automatically advance to the next page without participants having to click the Advance button, we found a JavaScript routine posted on the Qualtrics support forum. After participants clicked the cursor-centering button, the JavaScript routine automatically advanced to the next page before participants had the opportunity to click the Advance button.

After the JavaScript routine in the cursor-centering button triggered an auto-advance, a page containing the experimental stimulus and responses appeared. We generated the stimulus displays in PowerPoint then exported them into JPEG files that we uploaded to our Qualtrics image library. Screenshot examples appear in Figure 1. At the top of the page was a target item that was one of four capitalized color words (“Red,” “Green,” “Blue,” or “Yellow”) that matched the pixel color of the respective word against a black background. Below the target were four radio buttons, each of which was labelled with either a red, green, blue, or yellow patch of color (Stroop condition), or the word “Red,” “Green,” “Blue,” or “Yellow” written in black pixels (reverse Stroop condition). In the Stroop condition, participants were instructed to select the color patch that matched the target color, and in the reverse Stroop condition, participants were instructed to select the word that matched the target word. When participants selected the correct response, the JavaScript routine automatically advanced to the next trial’s cursor centering page. If participants clicked the Advance button instead of one of the four response buttons, Qualtrics registered that response as incorrect. In that case, and when participants actually did select an incorrect response, the script presented a message saying “The answer you provided was incorrect. Please try to select the right answer.” To continue the experiment after receiving this admonition, participants needed to click the Advance button, which then advanced to the next trial’s cursor centering page. We intended this extra step for incorrect responses to provide an incentive to select the correct responses.

**Procedure**

The first three pages in the Qualtrics script contained an informed consent letter, a set of demographic questions, and instructions, in that order. The instructions page informed participants that they would view a series of target items, one at a time, and that they should report the target color (for participants assigned to the Stroop condition in the first block) or word (for participants assigned to the reverse Stroop condition in the first block). After participants finished the instructions, the script proceeded through a series of four practice trials that were excluded from analysis, followed by 32 experimental trials.

The 32 experimental trials in each block contained 16 displays in which the target color and word were congruent (e.g., the word “Red” written in red pixels) and 16 in which they were incongruent (e.g., the word “Red” written in green pixels). The congruent displays included four repetitions of each of the four words “Red,” “Green,” “Blue,” and “Yellow,” written in pixels that matched the word’s meaning. The incongruent displays included two repetitions of each of the following eight word-color combinations: the word “Red” written in green pixels, “Red” in blue pixels, “Green” in red pixels, “Green” in yellow pixels, “Blue” in red pixels, “Blue” in yellow pixels, “Yellow” in green pixels, and “Yellow” in blue pixels. The target displays were the
same for both the Stroop and the reverse Stroop conditions, with the primary difference between the two conditions being that the response buttons were labelled with color patches in the Stroop condition or words written in black in the reverse Stroop condition. Thus, for example, when presented with an incongruent target such as the word “Red” in green pixels, the correct response in the Stroop condition would have been the green color patch, and the correct response in the reverse Stroop condition would have been the word “Red.” Given these 16 congruent displays and 16 incongruent displays, each of the four responses was correct eight times in each block. The 32 displays were presented in random order.

Each stimulus display page included a timer that was invisible to the participants. Timers in Qualtrics scripts record four pieces of data pertaining to the page where they are placed. For the parameters First click, Last click, and Page submit, the clock starts ticking when the page appears, and ends with the participant’s first mouse click on the page, the participant’s last click on the page, or when the page advances, respectively. The parameter Click count represents the number of times the participant clicked anywhere on the page before it advanced to the next page. In most Experiment 1a trials, participants clicked just once on the stimulus display page, which was when they clicked one of the response buttons. In trials with just one click, First click and Last click had an identical value, and Page submit was a few milliseconds longer, with the delay representing how long the JavaScript routine took to advance the page after a response button was selected. Because each page with a timer only had one question (asking participants to report the target color or word), in trials with just one click, we interpreted First click as the RT for that trial. Trials with more than one click were not as readily interpretable as those with just one click, as we will describe in the Results section.

After participants completed the four practice trials and 32 experimental trials, the script informed participants that they had just completed the first half of the experiment, and that for the remainder of the experiment they should attend to the other target feature. After reading the instructions, participants proceeded through four practice trials followed by 32 experimental trials presented in random order, while attending to the target feature that they had not attended in the first block. Once both blocks were completed, each participant was presented with a debriefing page.

### Results
The data from three participants were excluded from analysis. One participant did not finish the experiment, and the other two provided incorrect responses for more than half of the trials in at least one of the four conditions. In addition to the three excluded participants, we also removed RTs from individual trials for three reasons. First, we removed trials with incorrect responses, such as when a participant clicked the button for a blue color patch when the target’s pixels were red. The second reason for removal concerns the trials in which the Click count was greater than one. As mentioned previously in the Procedure section, for trials with a Click count of one, we interpreted the timer’s First click parameter as the RT for that trial. We did not initially understand why any trial would have more than one click, but eventually we discovered that either a finger swipe on a phone or tablet device to reposition the display, or a mouse click anywhere on the page not on a response button would be counted as a click. Because there is no way to determine whether an extra click was attributable to a finger swipe event on a tablet device or an actual extra click on a computer, we decided to exclude from analysis any trials in which Click count was greater than one. We should mention here that, on some trials, the Click count was zero, which occurred when a participant clicked the Advance button rather than one of the response buttons. Such instances were rare, and as mentioned previously in the Method section, they were recorded as incorrect responses and thus were removed along with the other incorrect responses. After removing any trials with incorrect responses or more than one click, we calculated mean RTs for each participant in the Stroop condition and the reverse Stroop condition. Any RT that was more than three standard deviations above the mean for its participant and condition was removed as an outlier. The percentage of trials that were removed for each of these three reasons is presented in Table 1.

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>Percentages of Trials Excluded From Analysis in Experiment 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Congruent</td>
</tr>
<tr>
<td>Errors</td>
<td>Clicks</td>
</tr>
<tr>
<td>Identification</td>
<td>Stroop</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
</tr>
<tr>
<td>Localization</td>
<td>Stroop</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
</tr>
</tbody>
</table>
Participants’ mean RT after errors, extra clicks, and outliers had been removed (depicted in Figure 2) were submitted to a two-way ANOVA with congruency and attended feature as within-subjects factors. Responses were faster for congruent targets than incongruent, $F(1, 79) = 16.96, p < .001, \eta_p^2 = .18$, but RTs were not significantly different when the participants attended to the target color (Stroop) than when they attended to the target word (reverse Stroop), $F(1, 79) = 3.22, p = .08, \eta_p^2 = .04$. The interaction between congruency and attended feature, $F(1, 79) = 11.03, p = .001, \eta_p^2 = .12$, shows that the Stroop effect was different than the reverse Stroop effect. Simple effects analysis confirmed that the Stroop effect was significant, $F(1, 79) = 57.36, p < .001, \eta_p^2 = .42$, but the reverse Stroop effect was not, $F(1, 79) = 0.02, p = .88, \eta_p^2 = .003$, and the Stroop effect size ($\eta_p^2 = .42$) was larger than the reverse Stroop effect size ($\eta_p^2 = .003$).

**Discussion**

The results from the identification task replicated the classic Stroop asymmetry; the Stroop effect was larger than the reverse Stroop effect. Because the classic Stroop asymmetry is sufficiently reliable to be called a classic, it may seem unsurprising that we succeeded. However, an online version of a Stroop task introduces much variability in the apparatuses and experimental contexts that could have prevented a successful replication. By obtaining the classic effect, the identification task lays the groundwork to suggest that, if the inversion of the classic Stroop effect in Sobel et al. (2020) is replicable, our remote version of their localization task should be sufficiently powerful to replicate it.

**Experiment 1b: Localization Task With University Students**

**Method**

**Participants**

Our sample included 90 participants, 13 of whom identified as male, 75 as female, and two as nonbinary. Participants’ ages ranged from 18 to 55 with a mean of 21.98 and standard deviation of 6.07. There were 64 participants who described themselves as White, 17 as Black or African American, three as Hispanic, two as Asian, two as American Indian or Alaska native, and one preferred not to state a race. As in the identification task, all participants were exposed to a Stroop condition in one block and a reverse Stroop condition in the other block, and block order was counterbalanced across participants.

**Apparatus**

As in the identification task, participants accessed the Qualtrics script using their own personal devices through their Sona account.

**Stimuli**

The basic structure of each trial was the same as in the identification task: each trial began with a cursor centering page, followed by a page containing the stimulus display and response buttons, and finally an admonishment page when participants selected an incorrect response. Also, the same JavaScript routines triggered automatic advancement in the cursor centering and stimulus display pages. However, the stimulus displays and response labels were different than in the identification task.
Each display contained not just one word as in the identification task, but all four of the color words (“Red,” “Green,” “Blue,” and “Yellow”), arranged on an imaginary circle centered on a fixation mark, as can be seen in the example screenshots in Figure 3. All four words were either congruent or incongruent with their pixel color. The items were placed at 90-degree intervals from each other, in one of two arrangements; in one arrangement, the four items appeared at clockface positions 1 o’clock, 4 o’clock, 7 o’clock, and 10 o’clock, and in the other arrangement the items appeared at 2 o’clock, 5 o’clock, 8 o’clock, and 11 o’clock. Below the target display was a cue indicating what the participant should search for in that trial. In the Stroop condition, the cue was a color patch containing pixels that were either red, green, blue, or yellow, and in the reverse Stroop condition, the cue was one of the words “Red,” “Green,” “Blue,” or “Yellow” written in black pixels. Below the target display and cue were two response buttons; the button on the left was labelled “Left side” and the button on the right was labelled “Right side.”

Procedure
As in the identification task, the first three pages in the Qualtrics script contained an informed consent letter, demographic questions, and instructions, in that order. The instructions page informed participants that they would view a series of displays and cues, and that they should report whether the target was positioned on the left side or the right side of the fixation cross. After reading the instructions, participants proceeded through a series of four practice trials that were excluded from analysis, followed by 32 experimental trials.

The 32 experimental trials in each block contained 16 congruent displays and 16 incongruent displays; congruent targets were written with pixels that were congruent with their meaning, and incongruent displays included the same eight word-color target combinations as in the identification task: “Red” in green pixels, “Red” in blue pixels, “Green” in red pixels, “Green” in yellow pixels, “Blue” in red pixels, “Blue” in yellow pixels, “Yellow” in green pixels, and “Yellow” in blue pixels. In the 16 congruent displays and the 16 incongruent displays, each word and each color appeared twice in each of the eight clockface positions. The target appeared on the right side of the fixation cross in half of the trials, and on the left side in the other half of the trials. The 16 congruent trials and 16 incongruent trials were presented in random order. At the end of the first block, participants were informed that they had completed half of the experiment, and they should attend to the other target feature for the remainder of the experiment. After completing both blocks, participants were presented with a debriefing statement.

Results
The data from 11 participants were excluded from analysis because in at least one of the four conditions more than half of their trials had more than one click. This is in sharp contrast to the results in Experiment 1a, in which none of the participants had so many trials with extra clicks. In our attempt to figure out why so many participants had extra clicks, we tried, among other things, running the script on a phone positioned horizontally. Because the stimulus displays were larger in Experiment 1b than the identification displays in Experiment 1a, as we ran the simulation we needed to swipe each display upward to access the response buttons. Thus we believe that 11 participants carried out the script on a phone they held horizontally, forcing them to swipe the display just to access the response buttons before they could even select one. In addition to the 11 excluded participants, we also removed RTs from individual trials for the same three reasons as in Experiment 1a. The percentages of removed trials are presented in Table 1.

Participants’ mean RTs (depicted in Figure 2) were submitted to a two-way ANOVA with congruity and attended feature as within-subjects factors. Responses were faster for congruent targets than incongruent, $F(1, 78) = 66.39, p < .001, \eta^2 = .46$, and were faster when participants attended to the target color (Stroop) than when they attended to the target word (reverse Stroop), $F(1, 78) = 39.78, p < .001, \eta^2 = .34$. The significant interaction between congruity and attended feature, $F(1, 78) = 23.31, p < .001, \eta^2 = .23$, shows that the Stroop effect was different than the reverse Stroop effect. Simple effects analysis confirmed that both the Stroop effect, $F(1, 78) = 7.06, p = .010, \eta^2 = .08$, and the reverse Stroop effect, $F(1, 78) = 54.80, p < .001, \eta^2 = .41$, were significant, and the Stroop effect size ($\eta^2 = .08$) was smaller than the reverse Stroop effect size ($\eta^2 = .41$).

In Experiment 1a, the two-way interaction between congruity and attended feature indicated that the Stroop effect was larger than the reverse Stroop effect, but in Experiment 1b, the two-way interaction indicated that the Stroop effect was smaller than the reverse Stroop effect. This suggests...
that the three-way interaction between congruity, attended feature, and task (identification versus localization) should be significant. With that in mind, we submitted mean correct RTs from both Experiments 1a and 1b to a three-way ANOVA with congruity and attended feature as within-subjects factors, and task as a between-subjects factor. The results confirmed that the three-way interaction was significant, $F(1, 157) = 32.83, p < .001, \eta^2_p = .17$.

**Discussion**

The main effect of congruity in the localization task in Experiment 1b echoed the same effect from the identification task Experiment 1a, but unlike the identification task, the main effect of attended feature was significant in Experiment 1b but not Experiment 1a. The Experiment 1b localization task required participants to search through several items before they could select one item for further processing, whereas no selection was required for the Experiment 1a identification task because there was only one display item. The main effect of attended feature in Experiment 1b indicates that responses were faster when participants attended to the target color in the Stroop condition than when they attended to the target word in the reverse Stroop condition. This effect was also present in Sobel et al. (2020), who argued that selecting one of the search items is more efficient for a visual feature such as color than for a semantic feature such as a word’s meaning (Wolfe & Horowitz, 2004).

The effect that is more relevant to our first hypothesis is the three-way interaction between congruity, attended feature, and task, which confirms our first hypothesis that an identification task would yield the classic Stroop asymmetry, and that a localization task would invert the asymmetry. This outcome represents more than a simple replication of Sobel et al. (2020), because our online version of identification and localization tasks introduced variability into the apparatuses and experimental contexts that would be absent from an experiment run on a single computer in a laboratory setting.

Nevertheless, our sample of undergraduate students is more homogeneous along such dimensions as age and educational level than the general population (Hanel & Vione, 2016; Peterson & Merunka, 2014). Of course, a major reason for the common practice in psychology departments (including our own) of allowing undergraduate students to gain course credit for their participation in experiments is that researchers can compensate participants for their time and attention without any money changing hands. Alas, members of the general public cannot be compensated with class credit. To gather a more representative sample for Experiment 2, we acquired external funding through a Psi Chi Undergraduate Research Grant to provide monetary incentives to members of the general public.

**Experiment 2a: Identification Task With a Nationwide Sample**

**Method**

**Participants**

With funding from a Psi Chi Undergraduate Research Grant, we were able to pay Qualtrics to recruit participants for Experiments 2a and 2b. Although the quote from Qualtrics did not specify how much they paid each participant, a simple calculation can estimate an upper boundary. We paid Qualtrics $1,050 and received data from 305 participants, which amounts to $3.44 per person, minus any brokering fees retained by Qualtrics. Because Qualtrics reports the latitude and longitude of the device used to connect to the online script, we could determine each participant’s general location. There was at least one participant from the four corners of the contiguous United States: from San Diego, CA, in the southwest; Seattle, WA, in the northwest; Miami, FL, in the southeast; Boston, MA, in the northeast; and many places in between. Of the 161 participants in Experiment 2a, 53 identified as male, 107 as female, and one as nonbinary. There were 110 participants who described themselves as White, 22 as Black or African American, 14 as Hispanic, five as Asian, five as American Indian or Alaska Native, two as Native Hawaiian or Pacific Islander, one as Creole, and two preferred not to report a race. Participants’ ages ranged from 18 to 75 years with a mean of 38.43 and standard deviation of 14.88.

**Apparatus, Stimuli, and Procedure**

As in Experiment 1, participants used their own personal devices to access the script on Qualtrics.com, which was identical to the one used in Experiment 1.

**Results**

The data from 40 participants were excluded from analysis because for each of these participants more than half of the trials from at least one of the four conditions included an incorrect response (17 participants), extra clicks (19 participants), or a combination of errors and extra clicks (four participants). In addition, the data from another
participant was excluded because their mean RT was more than three standard deviations greater than the mean RT of the other participants.

We have little insight into why there was a much higher proportion of participants excluded from the general population sample than from the sample of university students in Experiment 1, so any explanation we offer would necessarily be speculative. One possibility might be that university students can more readily imagine themselves as the researcher carrying out the project in which they participate than members of the general population, and thus the sample of university students included a higher proportion of individuals who conscientiously attended to the stimuli. To be clear, we do not mean to suggest that every member of the sample from the general population paid less attention than every member of the sample of university students. Instead, we are arguing that the sample from the general population might have included more individuals who were less invested in the experiment than the sample of university students. Researchers who carry out projects online are commonly urged to include attention checks (e.g., “If you are paying attention select strongly disagree”) as a way to remove the data from inattentive participants (Newman et al., 2021), but our task has built-in attention checks in the form of errors and extra clicks. With that in mind, after excluding inattentive participants as described above, we consider the remaining data to represent participants who were paying sufficient attention. In any event, we sought grant funding to obtain a more heterogeneous sample than could be obtained from students at our university, and heterogeneity from students at our university, and heterogeneity is just what we got: demographic, geographic, and attentional. Perhaps the need to remove a relatively high proportion of participants is simply the cost of doing business with heterogeneous samples.

As in Experiment 1, we also removed RTs from individual trials that were incorrect, had more than one click, or were more than three standard deviations greater than the mean RTs for that participant and task. The percentage of trials that were removed for each of these three reasons is presented in Table 2.

Participants’ mean RTs (depicted in Figure 4) were submitted to a two-way ANOVA with congruity and attended feature as within-subjects factors. Responses were faster for congruent targets than incongruent, $F(1, 119) = 50.04, p < .001, \eta_p^2 = .30$, but RTs were not significantly different when the participants attended to the target color (Stroop) than when they attended to the target word (reverse Stroop), $F(1, 119) = 0.22, p = .64, \eta_p^2 = .002$. The interaction between congruity and attended feature, $F(1, 119) = 37.42, p < .001, \eta_p^2 = .24$, shows that the Stroop effect was different than the reverse Stroop effect. Simple effects analysis confirmed that the Stroop effect was significant, $F(1, 119) = 106.96, p < .001, \eta_p^2 = .47$, but the reverse Stroop effect was not, $F(1, 119) = 3.25, p = .07, \eta_p^2 = .03$, and the Stroop effect size ($\eta_p^2 = .47$) was larger than the reverse Stroop effect size ($\eta_p^2 = .03$).

**Discussion**

The results from Experiment 2a echoed those from Experiment 1a, with a significant main effect of congruity but not attended feature, and a significant interaction between congruity and attended feature. Further, the presence of a classic Stroop asymmetry, as indicated by the interaction effect, lends preliminary support to our second hypothesis that the results from a convenience sample of undergraduate students would generalize to a more representative sample. To see if the

<table>
<thead>
<tr>
<th>TABLE 2</th>
<th>Percentages of Trials Excluded From Analysis in Experiment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Congruent</td>
</tr>
<tr>
<td>Errors</td>
<td>Clicks</td>
</tr>
<tr>
<td>Identifi.</td>
<td>Stroop</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
</tr>
<tr>
<td>Localization</td>
<td>Stroop</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
</tr>
</tbody>
</table>

Mean Response Times for an Identification Task in Experiment 2a and a Localization Task in Experiment 2b

![FIGURE 4](chart.png)

Note: Error bars represent 95% confidence intervals (calculation based on Loftus & Masson, 1994).
results from a localization task would also support our second hypothesis, in Experiment 2b, a sample from the general population completed the same online script as the participants in Experiment 1b.

**Experiment 2b: Localization Task With a Nationwide Sample**

**Method**

**Participants**

With money from our Psi Chi grant, we paid Qualtrics to recruit 144 participants, who were even a bit more geographically widespread than in Experiment 2a; one participant accessed the script from a device in Fairbanks, AK. Our sample included 47 participants who identified as male, 96 as female, and one as nonbinary. There were 92 participants who described themselves as White, 28 as Black or African American, 14 as Hispanic, seven as Asian, two as Native Hawaiian or Pacific Islander, and one as American Indian or Alaska Native. Participants’ ages ranged from 19 to 79 years with a mean of 39.74 and a standard deviation of 14.91.

**Apparatus, Stimuli, and Procedure**

The apparatus was the same as in Experiment 2a. The stimuli and procedure were the same as in Experiment 1b.

**Results**

The data from 56 participants were excluded from analysis for having errors (36 participants), extra clicks (15 participants), or a combination of errors and extra clicks (five participants) in more than half of the trials from at least one of the four conditions. As mentioned in the Results section for Experiment 2a, a higher proportion of participants were excluded from the general population sample than the sample of university students for reasons which elude us. Nevertheless, we can identify a possible reason why a higher proportion of participants were excluded from Experiment 2b than Experiment 2a. In contrast to the identification task, the localization task in Experiment 2b required two mental processing stages: a selection stage followed by a decision stage (Smith & Sewell, 2013). In each localization trial, participants were presented with not just one item, but four, and they needed to select one of the items before they could decide whether it was the target. Apparently, the extra layer of processing in localization led to more errors than just the single decision stage in identification. The percentage of individual trials excluded from analysis due to error, extra clicks, or being an outlier, is presented in Table 2.

Participants’ mean RTs (depicted in Figure 4) were submitted to a two-way ANOVA with congruity and attended feature as within-subjects factors. Responses were faster for congruent targets than incongruent, $F(1, 87) = 57.01, p < .001, \eta_p^2 = .40$, and responses were faster when the participants attended to the target color (Stroop) than when they attended to the target word (reverse Stroop), $F(1, 87) = 23.86, p < .001, \eta_p^2 = .22$. The interaction between congruity and attended feature, $F(1, 87) = 21.44, p < .001, \eta_p^2 = .20$, shows that the Stroop effect was different than the reverse Stroop effect. Simple effects analysis confirmed that both the Stroop effect, $F(1, 87) = 18.86, p < .001, \eta_p^2 = .18$, and the reverse Stroop effect, $F(1, 87) = 48.76, p < .001, \eta_p^2 = .36$, were significant, but the Stroop effect size ($\eta_p^2 = .18$) was smaller than the reverse Stroop effect size ($\eta_p^2 = .36$).

As in Experiment 1, the Stroop effect was larger than the reverse Stroop effect for an identification task, but smaller than the reverse for localization, which suggests that the three-way interaction between congruity, attended feature, and task (identification versus localization) should be significant. And also as in Experiment 1, a three-way ANOVA with congruity and attended feature as within-subjects factors, task as a between-subjects factor confirmed that the three-way interaction was significant, $F(1, 206) = 51.83, p < .001, \eta_p^2 = .20$.

**Discussion**

The results from Experiment 2b echoed those from Experiment 1b, with significant main effects of both congruity and attended feature; as in Experiment 1b, the main effect of attended feature suggests that target selection is more efficient on the basis of color than word. Also, the inversion of the classic Stroop asymmetry resulting from manipulating the task between Experiments 2a and 2b, as indicated by the three-way interaction of congruity, attended feature, and task, fully supports our second hypothesis that the results from a convenience sample of undergraduate students would generalize to a more representative sample.

**General Discussion**

Replications of the Stroop effect are vastly more common in the Stroop literature than replications of the reverse Stroop effect (Blais & Besner, 2006; Díaz-Piedra et al., 2022). Of course, it is impossible to determine whether the relative scarcity of reported replications indicates that noone
is looking for the effect, or that everyone who looks for an effect fails to find one. Even so, the strength-of-association account (Blais & Besner, 2007) suggests that the latter explanation is more likely, primarily because the typical task in Stroop experiments relies on identification of the target features, which is more strongly associated with the verbal feature.

In contrast, many studies that report either a reverse Stroop effect by itself (Blais & Besner, 2007; Diaz-Piedra et al., 2022; Yamamoto et al., 2016) or larger reverse Stroop effect than Stroop (Durgin, 2000; Miller et al., 2016; Sobel et al., 2020; Song & Hakoda, 2015; Uleman & Reeves, 1971) diverge from the traditional identification by using a different task (e.g., scanning, matching, searching, or pointing at a gun) that is more strongly associated with visual than verbal processing. Sobel et al. (2020), and more recently, Diaz-Piedra et al. (2022) argued that the strength of association between the task demands and the attended feature is the key for eliciting a robust reverse Stroop effect. In fact, if visual tasks were as common as verbal tasks throughout the long history of Stroop research, perhaps the classic Stroop asymmetry would never have become a classic, because inverted Stroop asymmetries would have been observed just as often as the asymmetry that has become known as the classic.

In this project, we aimed to develop online versions of Stroop identification and localization tasks, in the hope that a classic Stroop asymmetry for the identification task would validate our script, and an inversion of the classic Stroop asymmetry would support the strength-of-association account. The three-way interaction between congruity, attended feature, and task in Experiment 1 confirmed our first hypothesis, and the three-way interaction in Experiment 2 confirmed our second hypothesis. It would be tempting to conclude from these results that experiments conducted online are just as reliable as those in a laboratory, and researchers can readily generalize their findings obtained from a convenience sample of undergraduate students to the overall population. Nevertheless, we must acknowledge limitations of our project that prevent us from drawing such bold conclusions.

Limitations
As we acknowledged in the Introduction, online experiments entail sources of variability in computer hardware, software, and ambient distractions that cannot be eliminated (Anwyl-Irvine et al., 2021). In addition, students often report that they can more easily pay attention to lectures when they attend class in person than when they connect remotely (Becker et al., 2020), and the same may be true of participation in experiments. These sources of variability, as well as others that we are failing to imagine, seem likely to bias online experiments toward type II error, insofar as the variability might conceal an actual difference between conditions. While this is a limitation of our experiment as well as any other online experiment, in the Conclusion we will emphasize the upside.

A second limitation concerns whether the results obtained from a convenience sample generalize to a sample that is more representative of the overall population. We found that the results from a convenience sample in Experiment 1 did indeed generalize to a more representative sample in Experiment 2, but this is inconsistent with previous research showing that the results obtained from a convenience sample diverge unpredictably from more representative samples (Hanel & Vione, 2016; Peterson & Merunka, 2014). Of course, one relevant difference between our study and theirs is that our tasks assessed basic perceptual and attentional processing, whereas their studies examined participants’ explicit attitudes. This suggests that certain kinds of findings, such as those that describe basic cognitive mechanisms of perception and attention, may more readily generalize from convenience samples of undergraduate students to the overall population than other kinds of findings, such as those that describe explicit attitudes. The second limitation of our study is that we studied just one kind of cognitive task; before researchers who study cognitive processing in convenience samples can be confident that their results generalize to the overall population, many other cognitive tasks need to be compared across convenience samples and representative samples.

Conclusion
In our description of the first limitation, we acknowledged that known and unknown sources of variability entailed by online experiments bias them toward type II errors. This suggests that online experiments may be too insensitive to detect subtle effects, but could uniquely suit them for replication studies. If an effect can overcome a bias for a type II error, that would provide persuasive evidence in support of the hypothesis. Furthermore, our results combine with results from other recent studies (Diaz-Piedra et al., 2022; Sobel et al., 2020)
to provide converging evidence supporting the strength-of-association account of the Stroop asymmetry. Our second limitation recognized that our examination of the generalizability from convenience samples only looked at one kind of cognitive task, but this is nevertheless an important first step. In summary, we believe that our study makes a valuable contribution to the current understanding of the classic Stroop asymmetry.

References

Bausell, R. B., & Li, Y. F. (2002). Strength-of-association account of the Stroop asymmetry. In summary, we believe that our study makes a valuable contribution to the current understanding of the classic Stroop asymmetry.

Author Note. Mary E. Smith is now at the Department of Psychology, University of Memphis. This research was supported by an Undergraduate Research Grant from Psi Chi at the University of Central Arkansas. We have no known conflict of interest to disclose.
Turn Your Compassion Into a Career in Mental Health

Our online degree programs feature relationship-focused curriculum, innovative faculty, and lead to licensure.

alliant.edu/psychology
Call for Doctoral-Level Reviewers

*Psi Chi Journal* is seeking reviewers interested in providing constructive feedback on our authors’ original empirical research. To increase the journal’s scope and relevance, diverse people of varied racial, ethnic, gender identity, sexual orientation, religious, and social class backgrounds are welcomed. If you have a doctoral degree in psychology or a related field, please join us.

**TO SUPPORT OUR EXCEPTIONAL REVIEWERS**
- Our online portal allows you to submit specific subject areas that you feel comfortable with reviewing.
- At any time, you can set unavailability dates so that we will know when not to contact you.
- With each request to review, you will receive basic instructions and a template review form, which you can use in order to save you time during your review.

**TO BECOME A REVIEWER, EMAIL**
steve.rouse@psichi.org

---

An Eye on Graduate School

PsI Chi’s digital anthology brings together our very best advice about applying to graduate school—advice accumulated from 25+ experts in over 20+ years of *Eye on Psi Chi* magazine issues.

In nineteen chapters, this eBook will help you navigate the seven primary steps that are vital to your acceptance at the graduate program of your choice.

- Preparing for and Selecting a Graduate Program
- The Application Process
- Preparing for the GRE
- Soliciting Letters of Recommendation
- Writing the Personal Statement
- Interviewing
- Choosing a Program and Succeeding in Graduate School

**Download Today at Store.PsiChi.org**
Log in as a Psi Chi Member to receive a discount!

---

**Guidance Through a Successful Application**

Kindle Edition Available Now!
Publish Your Research in *Psi Chi Journal*

Undergraduate, graduate, and faculty submissions are welcome year round. Only one author (either first author or coauthor) is required to be a Psi Chi member. All submissions are free. Reasons to submit include

- a unique, doctoral-level, peer-review process
- indexing in PsycINFO, EBSCO, and Crossref databases
- free access of all articles at psichi.org
- our efficient online submissions portal

View Submission Guidelines and submit your research at [www.psichi.org/?page=JN_Submissions](http://www.psichi.org/?page=JN_Submissions)

---

Become a Journal Reviewer

Doctoral-level faculty in psychology and related fields who are passionate about educating others on conducting and reporting quality empirical research are invited to become reviewers for *Psi Chi Journal*. Our editorial team is uniquely dedicated to mentorship and promoting professional development of our authors—Please join us!

To become a reviewer, visit [www.psichi.org/page/JN_BecomeAReviewer](http://www.psichi.org/page/JN_BecomeAReviewer)

---

Resources for Student Research

Looking for solid examples of student manuscripts and educational editorials about conducting psychological research? Download as many free articles to share in your classrooms as you would like.

Search past issues, or articles by subject area or author at [www.psichi.org/journal_past](http://www.psichi.org/journal_past)

---

Learn About Psi Chi

Psi Chi is the International Honor Society in Psychology. Membership is primarily open to undergraduates, graduate students, transfer students, full-time and part-time faculty members, and alumni.

See membership benefits and a link to apply at [www.psichi.org/page/member_benefits](http://www.psichi.org/page/member_benefits)

---

Register an account:
[http://pcj.msubmit.net/cgi-bin/main.plex](http://pcj.msubmit.net/cgi-bin/main.plex)