Title

Revisiting the Inverted-U: Congruency Tasks Reveal Divergent Developmental Trajectories

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 44(44)

Authors

Erb, Christopher D.
Germine, Laura
Hartshorne, Joshua K

Publication Date

2022

Revisiting the Inverted-U: Congruency Tasks Reveal Divergent Developmental Trajectories

Christopher D. Erb (christopher.erb@auckland.ac.nz)
School of Psychology, University of Auckland, 23 Symonds Street
Auckland, 1010 New Zealand

Laura Germine (lgermine@mclean.harvard.edu)
Institute for Technology in Psychiatry, McLean Hospital, 115 Mill Street
Belmont, MA 02478 USA

Joshua K. Hartshorne (joshua.hartshorne@bc.edu)
Department of Psychology and Neuroscience, Boston College, 275 Beacon Street
Chestnut Hill, MA 02467 USA

Abstract

The Simon, Stroop, and flanker tasks are commonly used to investigate cognitive control. However, it remains unclear whether these three tasks in fact measure the same cognitive abilities and in the same proportion. We take a developmental approach to this question: if the Simon, Stroop, and flanker tasks all roughly measure the same capacity, they should show similar patterns of age-related change. This is difficult to ascertain from prior developmental research, which typically relies on small sample sizes, large age bins, and procedures that complicate direct comparisons of performance between the tasks. We present data from two massive online studies: Study 1 included 9,642 native English speakers between 10 and 80 years of age who completed two-alternative forced-choice versions of the Simon and Stroop tasks, and Study 2 included 13,448 English speakers between 10 and 79 years of age who completed a two-alternative forced-choice version of the flanker task. To minimize the contribution of feature-integration effects on performance, analysis was restricted to response switch trials. Of the three tasks, only the flanker task revealed an inverted U-shaped developmental trajectory, with performance improving until approximately 24 years of age and declining starting around 39 years of age. In contrast, performance on the Simon task improved until around 34 years of age but did not reveal significant age-related declines in later adulthood. Accuracy in the Stroop task revealed age-related improvements until approximately 21 years of age, but these improvements coincided with age-related declines in response time performance during the same period. Although the Simon and Stroop tasks are commonly interpreted to target similar underlying processes, we observed near zero correlations between the congruency effects observed in each task in terms of both accuracy and response time. We discuss these results in light of recent debates regarding the suitability of these tasks for assessing developmental and individual differences in cognitive control.

Keywords: cognitive control; flanker task; lifespan development; Simon task; Stroop task

Introduction

The development of cognitive control is commonly described as following an inverted U-shaped trajectory, with pronounced gains observed between childhood and early adulthood, and subsequent declines observed across late adulthood (e.g., Dempster, 1992; Zelazo, Craik, & Booth, 2004). Although this characterization of cognitive control's development likely captures the general trends observed across a range of tasks, many of the details remain unclear, such as when cognitive control reaches its peak and whether different tasks show similar age-related declines in performance. Answering this question is a prerequisite for understanding the cognitive mechanisms and neural underpinnings of cognitive control development.

There are two main obstacles to answering such questions. First, there is increasing debate in the field concerning the extent to which the tasks commonly used to assess cognitive control tap into similar underlying processes (e.g., Draheim et al., 2019; Rey-Mermet, Gade, & Oberauer, 2018). This presents a practical problem (how to synthesize results across tests) and a theoretical problem (how to conceptualize what the tests are measuring). Second, most studies have taken an extremely coarse-grained approach to studying development, comparing at most a handful of age groups, leaving adolescence undersampled and adulthood nearly entirely unsampled. An admirably large study by Gathercole et al. (2014) illustrates the issue: they tested 557 participants from 7 age groups averaging 3, 4, 5, 8, 15, 25, and 67 years of age on a Simon task. Although admirable in its comparative precision, the data are necessarily consistent with many different age-related curves. Moreover, the second problem compounds the first: the low precision of measurement confounds any comparison of age-related changes for different cognitive control tasks.

Measuring Control with Congruency Tasks

The current study addresses the challenges outlined above by comparing finely measured age-related changes in performance on three congruency tasks commonly used to assess developmental and individual differences in cognitive control: the Simon task (Simon, 1969), the Stroop task (Stroop, 1935), and the Eriksen flanker task (Eriksen & Eriksen, 1974). All three tasks involve performing some task (e.g., indicating whether an arrow points to the left or right) under *congruent* conditions where some task-extraneous cue reinforces the target response (e.g., surrounding distractor arrows also on the screen point the same direction as the target) or *incongruent* conditions (e.g., the surrounding distractor arrows point in the opposite direction as the target;

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

see Methods for a detailed description of each congruency task). The *congruency effect* is calculated by subtracting each participant's average performance on congruent trials from their average performance on incongruent trials (e.g., $RT_{Incongruent} - RT_{Congruent}$). In the limit, individuals who are able to focus perfectly on the task should hypothetically show a congruency effect of 0.

Data from the Simon task generally suggest an inverted U-shaped development, with performance increasing between childhood and early adulthood and subsequently decreasing in later adulthood (Davidson et al., 2006; Erb & Marcovitch, 2019; Gathercole et al., 2014; Van der Lubbe & Verleger, 2002; Vu & Proctor, 2008). However, because the majority of previous studies used small samples and investigated only a few age groups, it is difficult to say with much precision where the top of the U is – particularly since a number of studies fail to observe one (Christ et al., 2001; Sheridan et al., 2014). The most compelling evidence is from the aforementioned Gathercole et al. (2014) study, which suggested a peak somewhere between the ages of 15 and 67.

With regard to the Stroop task, the literature is less clear. Multiple studies have reported age-related improvements in Stroop performance across childhood or between childhood and early adulthood (Comalli et al., 1962; Ikeda et al., 2011; Leon-Carrion, García-Orza, & Pérez-Santamaría, 2004; Schiller, 1966). However, the version of the Stroop task used in these studies differs substantially from the versions commonly used in contemporary research. For instance, none of the studies featured congruent stimuli in which both the color of the text and the meaning of the word cued the same response. Consequently, it is currently unclear how the congruency effect changes between childhood and adulthood in the Stroop task and whether the age-related changes observed in the Stroop task correspond to those observed in the Simon task.

At the other end of the lifespan, many Stroop studies have reported age-related increases in congruency effects between early adulthood and late adulthood (e.g., Cohn, Dustman, & Bradford, 1984; Comalli, Wapner, & Werner, 1962). Verhaeghen and De Meersman (1998) argued that such increases are better explained in terms of age-related reductions in processing speed rather than changes in cognitive control (for an updated review, see Verhaeghen, 2011). A recent massive online study by LaPlume et al. (2021) featuring over 40,000 participants aged 18 to 90 failed to observe evidence of age-related declines in interference processing (computed by evaluating performance on incongruent trials after controlling for processing speed). As noted by the authors, this finding is consistent with recent meta-analytic work suggesting that age-related declines in performance on congruency tasks are not observed if individual differences in processing speed are taken into account (Rey-Mermet & Gade, 2018). However, the authors also note that they may have failed to observe an age-related decline in performance because they used a number naming version of the Stroop task that is easier than the standard color-word version (citing Bugg et al., 2007).

The flanker task has featured prominently in research investigating age-related differences in cognitive control and has been standardized as a measure of inhibitory control by the NIH (Zelazo et al., 2014). Results from the task generally indicate that performance follows an inverted U-shaped trajectory, with peak performance occurring in early or middle adulthood followed by a decline in performance beginning in late adulthood (Erb, Touron, & Marcovitch, 2020; Luna et al., 2004; Waszak, Li, & Hommel, 2010). Waszak et al. (2010), for example, evaluated data from 263 participants ranging in age from 6 to 88 years of age and found that conflict resolution in the flanker task peaked between 35 and 42 years of age, with the effect of conflict increasing after the age of 67.

In summary, previous investigations of the Simon, Stroop, and flanker tasks present mixed evidence that the tasks follow an inverted U-shaped developmental trajectory. Discrepancies in the literature likely reflect a range of factors, including task differences (e.g., Bugg et al., 2007), differences in how conflict effects were computed (e.g., Christ et al., 2001), and the extent to which individual differences in processing speed were taken into account (e.g., Rey-Mermet & Gade, 2018). Further, recent work exploring trial sequence effects within congruency tasks indicates that performance on the tasks can reflect a range of factors that are often not taken into account when evaluating developmental differences, including contingency learning effects and feature-integration effects stemming from the repetition of stimulus and/or response features (for a review, see Braem et al., 2019). Consequently, it is possible that many of the developmental differences reported in the literature reflect the contribution of such effects to differing degrees depending on the design of the task (Erb & Marcovitch, 2018).

Although congruency tasks are commonly interpreted to tap into common cognitive processes relating to attentional control and inhibition, studies comparing congruency effects collected from the same group of participants across different tasks often reveal low correlations. For example, Draheim et al. (2019) found that the response time congruency effects observed in Stroop and flanker tasks typically correlate at approximately $r = .10$. As noted by the researchers, this likely reflects issues with the reliability of congruency tasks stemming in part from the calculation of difference scores in the presence of speed-accuracy trade-off effects. Such differences may also reflect task-specific abilities. For instance, Protopapas et al. (2007) found that reading expertise was negatively related to the Stroop effect in 7th graders, with higher reading scores corresponding to smaller congruency effects. Thus, the existing literature suggests that performance across different congruency tasks is unlikely to be strongly correlated, particularly in standard versions of the tasks that do not control for speed-accuracy trade-offs (Draheim et al., 2020). However, it is currently unclear whether these tasks might follow similar developmental trajectories despite not presenting strongly correlated effects within individuals.

Current Studies

The preceding discussion raised three central questions concerning how cognitive control, as assessed by the Simon, Stroop, and Eriksen flanker tasks, changes across the lifespan: (1) *When is peak performance observed in each task?* (2) *Do the tasks present similar developmental trajectories?* And, (3) *Is performance correlated across congruency tasks?* Addressing these questions is particularly important in light of recent work that has used large sample sizes and small age bins to demonstrate considerable heterogeneity in when other fundamental cognitive abilities like intelligence and memory peak (Hartshorne & Germine, 2015).

To address these issues, we collected two massive datasets using online Citizen Science samples. Unlike studies that recruit participants with offers of course credit or cash, Citizen Science relies on intrinsically-motivating tasks. Critically, data quality matches or exceeds what can be obtained in the lab, perhaps precisely because subjects are more likely to be engaged (Germine et al., 2012; Meyerson & Tyron, 2003; Ye et al., 2017).

Critically, Citizen Science experiments can recruit large samples covering much of the lifespan, allowing precise measurement of age-related change. Not surprisingly, such studies have become increasingly influential in studies of lifespan development, where they have revealed heretofore unsuspected developmental patterns, radically changing the theoretical landscape (Hartshorne, 2020; Hartshorne & Germine, 2015; LaPlume et al., 2021).

We therefore present data from two massive online studies. Study 1 includes 9,585 native English speakers between 10 and 77 years of age who completed two-alternative forced-choice versions of the Simon and Stroop tasks. Study 2 includes 13,449 English speakers between 10 and 79 years of age who completed a two-alternative forced-choice version of the flanker task through TestMyBrain.org. Data from Study 2 will therefore be used to address research questions (1) and (2).

Methods

Participants

Study 1. 9,642 participants the Stroop and/or Simon tasks: 9,576 participants completed the Stroop task (mean age in years = 31, SE = 16; 3,119 male, 6,457 female), 9,585 participants completed the Simon task (mean age in years = 31, SE = 16; 3,114 male, 6,471 female), and 9,460 participants completed both. Additional participants were excluded for being repeat participants (N = 377), reporting not-corrected-to-normal vision (N = 469), reporting dyslexia or a neurological disorder (N = 1,073), or software error (Stroop = 163, Simon = 133). Finally, we excluded 81 subjects who claimed to be less than 10 or more than 80 years old. This cutoff was determined by finding the oldest age for which we had at least 10 subjects and the youngest age over 5 for which we had at least 10 subjects (there tends to be an over-abundance of 0 and 1-year-olds, reflecting participants who did not wish to fill out the demographics, choosing a young age is the fastest way of advancing to the task on the online platform).

Study 2. A total of 13,449 English speakers completed the experiment. Additional participants were excluded for being repeat participants (N = 1,915) or software error (N = 1,007) (no info was collected on vision or neurological disorders). Finally, we excluded 84 participants who claimed to be less than 10 or more than 79 years old. As with Study 1, this cutoff was determined by finding the oldest age for which we had at least 10 participants and then the youngest age over 5 for which we had at least 10 participants.

Procedure

Study 1. Participants completed 2AFC versions of the Simon task and the Stroop task in a randomized order. Each task consisted of 14 congruent trials and 14 incongruent trials, presented in a randomly intermixed order that was held constant across tasks and individuals. On each trial in the Simon task, participants were presented with a left-facing arrow or a right-facing arrow at either the left or right side of their display. Participants were instructed to press the "w" key for arrows pointing to the left and the "o" key for arrows pointing to the right, regardless of which side of the screen the arrow was presented on. This version of the Simon task is sometimes referred to as the "spatial Stroop" task (Lu & Proctor, 1995; Williams et al., 2007), as the correct response is specified directly by the stimulus as opposed to versions of the task that use non-directional stimuli (e.g., an "O" vs. an "X" or a blue square vs. a red square). On each trial in the Stroop task, participants were presented with the word "orange" or "white" in either orange or white text presented against a grey background. Participants were instructed to press the "w" key if the word appeared in white text and the "o" key if the word appeared in orange text. For both tasks, the ITI was 200 ms and no feedback was given. Participants had as much time to respond as they wished. Each task began with a 4-item practice block. If subjects made any errors, they were informed of this and the instructions and practice were repeated.

Study 2. Participants completed a 2AFC version of the Eriksen flanker task in which a stimulus array consisting of 5 arrows appeared on each trial. On congruent trials, each of the arrows cued the same response (e.g., "< < < < <"). On incongruent trials, the centrally presented target arrow cued a different response than the surrounding distractors (e.g., "> > < > >"). The distractors were presented before the target for 100 ms. The distractors and target were then presented together for 50 ms, followed by a blank screen for 70 ms. A fixation cross then remained on the screen until a response was given or 3,000 ms elapsed. The ITI was set at 1,000 ms. Participants responded by pressing the "x" key (on keyboards) or a left-side software button (on touch-screens) for left-facing target arrows and the "m" key (on keyboards) or a right-facing software button (on touch-screens) for right-facing arrows. The task consisted of 96 trials, with congruent/incongruent and left/right responses fully crossed.

Response types were in the same order for all participants. Trials in which no response was provided within the 3,000 ms limit were marked as inaccurate (0.22% of all trials; 2.0% of incorrect trials). The task began with two 4-item blocks of practice: the first at a slower speed and the second at the true speed. Feedback was given after each response in the practice, and incorrect responses prompted that item to repeat.

Results

Data Processing

To allow for warm-up, the first trial of each block was excluded (1 trial each for Stroop and Simon; 3 for flanker). To control for post-error performance adjustments (e.g., Danielmeier & Ullsperger, 2011), all inaccurate trials and trials following an inaccurate trial were excluded from analysis. This resulted in the exclusion of 0.14% of Stroop trials, 0.08% of Simon trials, and 0.30% of flanker trials. Additionally, all responses that were faster than 100 ms or slower than 1,000 ms were excluded from analysis, resulting in the further exclusion of 0.16% of Stroop trials, 0.03% of Simon trials, and 0.03% of flanker trials (the latter number includes trials that timed out at 3,000 ms). To improve linearity, response times were transformed with the natural logarithm.

Minimizing Feature-Integration Confounds

To minimize the contribution of feature-integration effects, data analysis was restricted to response alternation trials. This ensured that trials featuring a repetition of the stimulus and response presented on the preceding trial were excluded from analysis, as such repetitions have been found to facilitate performance. Similarly, this ensured that trials featuring the same response but a different stimulus than the preceding trial were excluded from analysis, as this subset of trials is proposed to generate conflict during stimulus-response binding (Erb & Marcovitch, 2018; Hommel, 2004; Nieuwenhuis et al., 2006). This resulted in the inclusion of 28 trials from the Simon task, 28 trials from the Stroop task, and 43 trials from the flanker task after the first trial of each block was excluded.

Internal Reliability

To evaluate the internal reliability of our congruency effect measures, we estimated Chronbach's alpha for the average split-half correlation, averaged over 100 random splits. Reliability was marginally adequate for accuracy in the Simon task (0.80, CI: 0.77, 0.83) and flanker task (0.65, CI: 0.62, 0.67), but comparatively low in the Stroop task (0.47, CI: 0.45, 0.50). Log-transformed response times revealed low reliability for the flanker task (0.31, CI: 0.25, 0.36) and essentially no reliability for the Simon task (0.04, CI: -.22, 0.08) or the Stroop task (0.02; CI: -0.05, 0.08). These results echo concerns about the reliability of standard congruency tasks for investigating developmental and individual differences (e.g., Draheim et al., 2019) and indicate that the following result sections should be interpreted with caution, particularly with regard to the response time findings.

Peak Performance and Developmental Trajectories

To evaluate age-related changes in the size of the congruency effect observed in accuracy and response times for each task, we performed separate Bayesian thin plate spline regressions for each task using the brms package (Bürkner, 2017; Carpenter et al., 2017).[1] The resulting curves are shown in Figure 1 (accuracy) and Figure 2 (response times). Given that reliability analyses indicated that the congruency effects observed in response time were very low, we again emphasize that the response time results should be interpreted with caution.

To quantify how the curves change with age, we calculated the local slope at every 1/100th of a year. This was repeated for each sampled curve, allowing us to estimate uncertainty. Intervals where the slope's 95% credible interval excluded 0 are intervals of significant developmental change. Note that this method inherently corrects for multiple comparisons (Simpson, 2016).

Simon Task. The accuracy congruency effect decreased from the youngest age group until approximately 34 years of age, with a brief hiatus in the early 20s, with no further significant age-related changes observed. Response times revealed no significant age-related changes in the size of the congruency effect.

Stroop Task. The accuracy congruency effect decreased from the youngest age group until approximately 21 years of age. Although the size of the effect began to grow again in later life, this numerical trend did not reach statistical significance. Interestingly, the congruency effect observed in response times increased until approximately 21 years of age. Although numerically there appears to be a decrease in the size of the congruency effect observed in response times starting in middle age, this trend did not reach significance.

Flanker Task. The accuracy congruency effect decreased from the youngest age group until approximately 24 years of age. The accuracy congruency effect then started to increase at approximately 39 years of age and continued this trend through the oldest age group tested. The congruency effect observed in response times increased from approximately 41 to 57 years of age.

---

[1] We used 6 chains of 2,500 iterations each, with adapt delta of .99 and max_treedepth of 15. Priors on regression coefficients were set to N(0, 0.5).

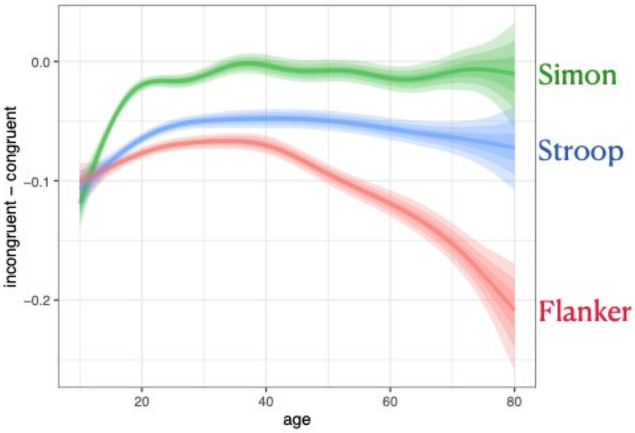

Figure 1: Accuracy congruency effects observed in the Simon task (green), Stroop task (blue), and Eriksen flanker task (red) as a function of age in years. Shading denotes 50%, 80%, and 95% confidence intervals.

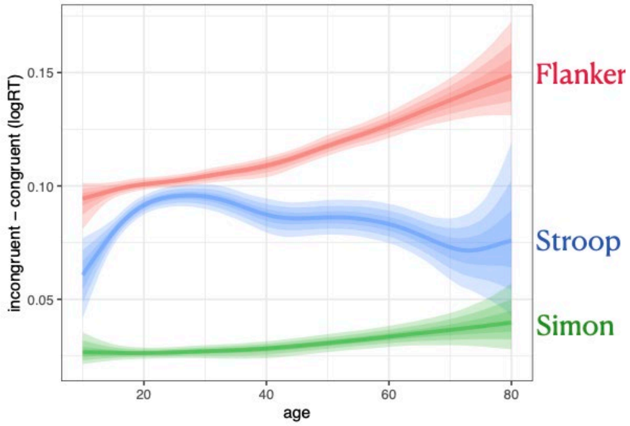

Figure 2: Response time congruency effects observed in Simon task (green), Stroop task (blue), and Eriksen flanker task (red) as a function of age in years. Shading denotes 50%, 80%, and 95% confidence intervals.

Correlations Between Simon and Stroop Effects

The correlation between congruency effects for the Simon and Stroop tasks was near zero for both dependent measures, despite the two tasks having identical trial structure and closely-matched procedures. As illustrated in Figure 3, the correlations were largely unchanged across the age range investigated.

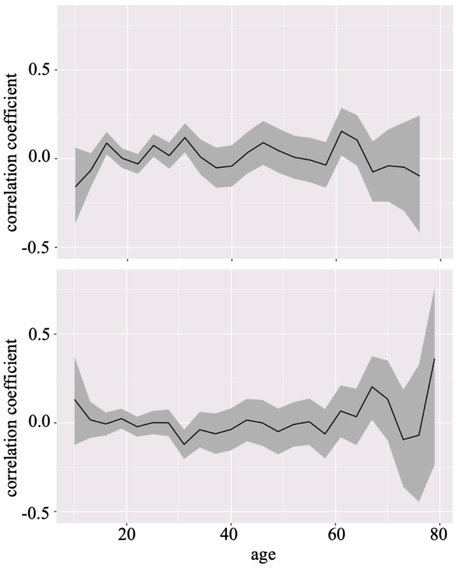

Figure 3: Correlations between the Simon and Stroop congruency effects observed in accuracy (top) and response time (bottom) as a function age. Age groups presented in 3-year bins. Shading denotes 95% confidence intervals.

Discussion

Cognitive control is commonly proposed to follow an inverted U-shaped developmental trajectory. However, a detailed understanding of cognitive control's development has proven elusive. This point is well illustrated by a recent review of research investigating the effect of aging on flanker performance in the Attention Network Task by Verissimo et al. (2022). Of the thirteen reviewed studies, eight reported an age-related decrease in control in later adulthood, seven reported no age-related changes, and four reported age-related improvements. As noted in the introduction, similar discrepancies have been observed in research with the Simon and Stroop tasks. These discrepancies likely reflect a range of factors, including issues with reliability (Draheim et al., 2019), small correlations among congruency effects (Rey-Mermet et al., 2018), and the prevalence of unacknowledged confounds (Braem et al., 2019).

The current study focused on yet another factor that is likely to contribute to inconsistencies in the literature: the use of small sample sizes, with large age bins, and categorical designs that compare a limited number of age groups. To address this limitation, we conducted two massive online studies to track the development of cognitive control in the Simon, Stroop, and Eriksen flanker tasks.

In contrast to the notion that these tasks exhibit similar developmental trajectories, we observed pronounced differences across the tasks. For instance, peak performance occurred at approximately 24 years of age in the flanker task and at approximately 34 years of age in the Simon task. Although the Stroop task revealed age-related improvements in accuracy until approximately 21 years of age, these improvements coincided with age-related *decreases* in response time performance during the same period.

The tasks also revealed different effects of aging across adulthood. Neither the Simon task nor the Stroop task revealed significant age-related declines in performance between early and late adulthood, with response times in the Stroop task showing a non-significant trend of age-related *improvements* in older adulthood. Only the flanker task showed strong evidence of an inverted U-shaped trajectory, with accuracy revealing age-related declines in performance starting at approximately 39 years of age and continuing until the oldest age group tested. Response times in the flanker task also revealed significant age-related declines in performance between approximately 41 and 57 years of age.

The lack of an inverted U-shaped developmental trajectory in Stroop performance is consistent with a recent massive online study by LaPlume et al. (2021) that failed to observe significant age-related declines in interference processing in a number-naming version of the Stroop task. Given that we also failed to observe such declines with the classic color-word version of the task, our results suggest that the results of LaPlume et al. (2021) were not specific to their version of the task.

The developmental trajectory observed in the flanker task is generally consistent with the trajectory observed by Waszak et al. (2010) in a sample of 263 participants ranging from 6 to 88 years of age. Waszak et al. focused on the conflict effect observed in response times and also reported an inverted U-shaped trajectory in performance. However, peak performance occurred substantially later in their sample, between 35 and 42 years of age. Age-related declines also began later in the Waszak et al. study, with significant increases in the size of the conflict effect only starting after the age of 67. We suspect that these differences are attributable in part to the smaller sample size collected by Waszak et al., as visual inspection of their figures suggests that age-related declines in performance may have begun earlier but failed to reach statistical significance.

Neither our results nor those of Waszak et al. (2010) appear to be consistent with those of Verissimo et al. (2022), who reported age-related improvements in flanker performance (smaller congruency effects) between individuals in their late 50s and individuals in their mid to late 70s. This effect stemmed from age-related slowing having a more pronounced effect on congruent trials, resulting in the difference between congruent and incongruent response times decreasing. Interestingly, we observed the inverse effect in our flanker data, with more pronounced effects of aging in incongruent than congruent trials. For example, although accuracy on congruent trials decreases somewhat between middle and older adulthood, we observed much larger reductions in accuracy on incongruent trials during the same period. To the extent that age-related reductions in congruency effects are driven by more pronounced effects of aging on congruent trials, we would be hesitant to interpret such reductions as indexing improved cognitive control or conflict processing.

With regard to our correlational analyses, we failed to observe significant links between the congruency effects observed in the Simon and Stroop tasks. Recent diffusion modeling by Hedge et al. (2021) indicates that correlations between congruency effects are only weakly informative regarding the presence or absence of shared control processes even when other factors such as response caution and processing speed are taken into account. Consequently, the lack of significant correlations between the congruency effects observed in the current study may not indicate that the tasks fail to tap into common cognitive control processes. However, taken together with the divergent developmental trajectories observed across the tasks, our correlational results further underscore the need for caution when attempting to form general conclusions about the development of cognitive control by synthesizing results across different congruency tasks.

Finally, it is important to reiterate that internal reliability also varied substantially across the congruency effects observed in three tasks, with the congruency effects observed in accuracy revealing strong reliability in the Simon task (.80), marginal reliability in the flanker task (.65), and relatively low reliability in the Stroop task (.47). As a point of comparison, the flanker effect observed in response times in the aforementioned study by Waszak et al. (2010) was .44 (the reliability for the flanker effect observed in accuracy was not reported). Although LaPlume et al. (2021) did not report the reliability of the Stroop effect observed in their massive online study, previous research using the same version of the task reported a test-retest reliability of .83 for the response time Stroop effect observed in a sample of 76 individuals over 50 years of age (Troyer et al., 2014). Thus, our reliability metrics are generally comparable to the metrics reported in previous studies investigating the development of cognitive control.

Conclusion

Recent research has called into question the extent to which commonly used versions of the Simon, Stroop, and Eriksen flanker task can be used to measure the same cognitive abilities and in the same proportion (Draheim et al., 2019; Hedge et al., 2019; Rey-Mermet et al., 2018). We explored this question from a developmental perspective, reasoning that the tasks should show similar patterns of age-related change if they tap into shared cognitive abilities. Our results revealed markedly different developmental trajectories, with only the flanker task conforming to an inverted U-shape. These findings caution against using standard congruency tasks to draw general conclusions about the development of cognitive control and underscore the importance of developing more psychometrically rigorous measures (Draheim et al., 2019). Our findings also highlight the value of using large sample sizes and small age bins to test for heterogeneity in the development of cognitive abilities (Hartshorne & Germine, 2015).

References

Braem, S., Bugg, J. M., Schmidt, J. R., Crump, M. J., Weissman, D. H., Notebaert, W., & Egner, T. (2019). Measuring adaptive control in conflict tasks. *Trends in Cognitive Sciences*, *23*(9), 769-783.

Bugg, J. M., DeLosh, E. L., Davalos, D. B., & Davis, H. P. (2007). Age differences in Stroop interference: Contributions of general slowing and task-specific deficits. *Aging, Neuropsychology, and Cognition*, *14*(2), 155-167.

Bürkner, P. C. (2017). Advanced Bayesian multilevel modeling with the R package brms. arXiv preprint arXiv:1705.11123.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1-32.

Christ, S. E., White, D. A., Mandernach, T., & Keys, B. A. (2001). Inhibitory control across the life span. *Developmental Neuropsychology*, *20*(3), 653-669.

Cohn, N. B., Dustman, R. E., & Bradford, D. C. (1984). Age-related decrements in stroop color test performance. *Journal of Clinical Psychology*, *40*(5), 1244-1250.

Comalli Jr, P. E., Wapner, S., & Werner, H. (1962). Interference effects of Stroop color-word test in childhood, adulthood, and aging. *The Journal of Genetic Psychology*, *100*(1), 47-53.

Danielmeier, C., & Ullsperger, M. (2011). Post-error adjustments. *Frontiers in Psychology*, *2*, 233.

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*(11), 2037-2078.

Dempster, F. N. (1992). The rise and fall of the inhibitory mechanism: Toward a unified theory of cognitive development and aging. *Developmental Review*, *12*(1), 45-75.

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, *145*(5), 508-535.

Erb, C. D., & Marcovitch, S. (2018). Deconstructing the Gratton effect: Targeting dissociable trial sequence effects in children, pre-adolescents, and adults. *Cognition*, *179*, 150-162.

Erb, C. D., & Marcovitch, S. (2019). Tracking the within-trial, cross-trial, and developmental dynamics of cognitive control: Evidence from the Simon task. *Child Development*, *90*(6), e831-e848.

Erb, C. D., Touron, D. R., & Marcovitch, S. (2020). Tracking the dynamics of global and competitive inhibition in early and late adulthood: Evidence from the flanker task. *Psychology and Aging*, *35*(5), 729-743.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143-149.

Gathercole, V. C. M., Thomas, E. M., Kennedy, I., Prys, C., Young, N., Viñas-Guasch, N., ... & Jones, L. (2014). Does language dominance affect cognitive performance in bilinguals? Lifespan evidence from preschoolers through older adults on card sorting, Simon, and metalinguistic tasks. *Frontiers in Psychology*, 5, 11.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. P*sychonomic Bulletin & Review*, *19*, 847–857.

Hartshorne, J. K. (2020). How massive online experiments (MOEs) can illuminate critical and sensitive periods in development. *Current Opinion in Behavioral Sciences*, *36*, 135-143.

Hartshorne, J. K., & Germine, L. T. (2015). When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. *Psychological Science*, *26*(4), 433-443.

Hedge, C., Powell, G., Bompas, A., & Sumner, P. (2021). Strategy and processing speed eclipse individual differences in control ability in conflict tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, http://dx.doi.org/10.1037/xlm0001028

Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in Cognitive Sciences*, *8*(11), 494-500.

Ikeda, Y., Okuzumi, H., Kokubun, M., & Haishi, K. (2011). Age-related trends of interference control in school-age children and young adults in the Stroop color–word test. *Psychological Reports*, *108*(2), 577-584.

LaPlume, A. A., Anderson, N. D., McKetton, L., Levine, B., & Troyer, A. K. (2022). When I'm 64: Age-related variability in over 40,000 online cognitive test takers. *The Journals of Gerontology: Series B*, *77*(1), 104-117.

Leon-Carrion, J., García-Orza, J., & Pérez-Santamaría, F. J. (2004). Development of the inhibitory component of the executive functions in children and adolescents. *International Journal of Neuroscience*, *114*(10), 1291-1311.

Lu, C. H., & Proctor, R. W. (1995). The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. *Psychonomic Bulletin & Review*, *2*(2), 174-207.

Luna, B., Garver, K. E., Urban, T. A., Lazar, N. A., & Sweeney, J. A. (2004). Maturation of cognitive processes from late childhood to adulthood. *Child Development*, *75*(5), 1357-1372.

Meyerson, P., & Tryon, W. W. (2003). Validating Internet research: A test of the psychometric equivalence of Internet and in-person samples. *Behavior Research Methods, Instruments, & Computers*, *35*, 614–620.

Nieuwenhuis, S., Stins, J. F., Posthuma, D., Polderman, T. J., Boomsma, D. I., & de Geus, E. J. (2006). Accounting for sequential trial effects in the flanker task: Conflict adaptation or associative priming? *Memory & Cognition*, *34*(6), 1260–1272.

Protopapas, A., Archonti, A., & Skaloumbakas, C. (2007). Reading ability is negatively related to Stroop interference. *Cognitive Psychology*, *54*(3), 251-282.

Rey-Mermet, A., & Gade, M. (2018). Inhibition in aging: What is preserved? What declines? A meta-analysis. *Psychonomic Bulletin & Review*, *25*(5), 1695-1716.

Rey-Mermet, A., Gade, M., & Oberauer, K. (2018). Should we stop thinking about inhibition? Searching for individual and age differences in inhibition ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(4), 501-526.

Schiller, P. H. (1966). Developmental study of color-word interference. *Journal of Experimental Psychology*, *72*(1), 105-108.

Sheridan, M., Kharitonova, M., Martin, R. E., Chatterjee, A., & Gabrieli, J. D. (2014). Neural substrates of the development of cognitive control in children ages 5–10 years. *Journal of Cognitive Neuroscience*, *26*(8), 1840-1850.

Simon, J. R. (1969). Reactions toward the source of stimulation. *Journal of Experimental Psychology*, *81*(1), 174-176.

Simpson, G. (2016, December 15). Simultaneous intervals for smooths revisited: Correcting a silly mistake. *From the Bottom of the Heap.* https://fromthebottomoftheheap.net/2016/12/15/simultaneous-interval-revisited/

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643-662.

Troyer, A. K., Rowe, G., Murphy, K. J., Levine, B., Leach, L., & Hasher, L. (2014). Development and evaluation of a self-administered on-line test of memory and attention for middle-aged and older adults. *Frontiers in Aging Neuroscience*, *6*, 335.

Van der Lubbe, R. H., & Verleger, R. (2002). Aging and the Simon task. *Psychophysiology*, *39*(1), 100-110.

Verhaeghen, P. (2011). Aging and executive control: Reports of a demise greatly exaggerated. *Current Directions in Psychological Science*, *20*(3), 174-180.

Verhaeghen, P., & De Meersman, L. (1998). Aging and the Stroop effect: A meta-analysis. *Psychology and Aging*, *13*(1), 120-126.

Veríssimo, J., Verhaeghen, P., Goldman, N., Weinstein, M., & Ullman, M. T. (2021). Evidence that ageing yields improvements as well as declines across attention and executive functions. *Nature Human Behaviour*, *6*, 97-110.

Vu, K. P. L., & Proctor, R. W. (2008). Age differences in response selection for pure and mixed stimulus–response mappings and tasks. *Acta Psychologica*, *129*(1), 49-60.

Waszak, F., Li, S. C., & Hommel, B. (2010). The development of attentional networks: Cross-sectional findings from a life span sample. *Developmental Psychology*, *46*(2), 337-349.

Williams, B. R., Strauss, E. H., Hultsch, D. F., & Hunter, M. A. (2007). Reaction time inconsistency in a spatial Stroop task: Age-related differences through childhood and adulthood. *Aging, Neuropsychology, and Cognition*, *14*(4), 417-439.

Ye, T., Reinecke, K., & Robert Jr, L. P. (2017, February). Personalized feedback versus money: the effect on reliability of subjective data in online experimental platforms. In Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 343-346).

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., Conway, K. P., ... & Weintraub, S. (2014). NIH Toolbox Cognition Battery (CB): Validation of executive function measures in adults. *Journal of the International Neuropsychological Society*, *20*(6), 620-629.

Zelazo, P. D., Craik, F. I., & Booth, L. (2004). Executive function across the life span. *Acta Psychologica*, *115*(2-3), 167-183.