Journal of Neuroscience Methods 274 (2016) 81–93

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

The role of test-retest reliability in measuring individual and group differences in executive functioning

Kenneth R. Paap[a],*, Oliver Sawi[b]

[a] Department of Psychology, EP301, San Francisco State University, 1600 Holloway Avenue, San Francisco, CA 94132, USA
[b] Department of Psychology, University of Connecticut, Storrs, CT, USA

HIGHLIGHTS

- 12 measures of executive functioning were obtained from 4 tasks.
- Test-retest reliability was reported for each measure.
- Reliability of single mean measures was greater than for difference score measures.
- Implications for tests of executive functioning were discussed.

ARTICLE INFO

Article history:
Received 6 March 2016
Received in revised form
30 September 2016
Accepted 3 October 2016
Available online 6 October 2016

Keywords:
Executive functions
Reliability
Validity
Antisaccade
Flanker
Simon
Switching

ABSTRACT

Background: Studies testing for individual or group differences in executive functioning can be compromised by unknown test-retest reliability.
New method: Test-retest reliabilities across an interval of about one week were obtained from performance in the antisaccade, flanker, Simon, and color-shape switching tasks. There is a general trade-off between the greater reliability of single mean RT measures, and the greater process purity of measures based on contrasts between mean RTs in two conditions. The individual differences in RT model recently developed by Miller and Ulrich was used to evaluate the trade-off.
Results: Test-retest reliability was statistically significant for 11 of the 12 measures, but was of moderate size, at best, for the difference scores. The test-retest reliabilities for the Simon and flanker interference scores were lower than those for switching costs.
Comparison with existing methods: Standard practice evaluates the reliability of executive-functioning measures using split-half methods based on data obtained in a single day. Our test-retest measures of reliability are lower, especially for difference scores. These reliability measures must also take into account possible day effects that classical test theory assumes do not occur.
Conclusions: Measures based on single mean RTs tend to have acceptable levels of reliability and convergent validity, but are "impure" measures of specific executive functions. The individual differences in RT model shows that the impurity problem is worse than typically assumed. However, the "purer" measures based on difference scores have low convergent validity that is partly caused by deficiencies in test-retest reliability.

1. Introduction

1.1. Purpose

The purpose of this paper is to present cross-session and within-session correlations for a set of measures that are often assumed to reflect particular executive functions (EFs). Based on these functional interpretations researchers (e.g., Unsworth et al., 2014; Paap and Sawi, 2014) have explored the construct validity of some proposed EFs by looking for correlations or dissociations across measures. Our focus here is on the test-retest reliability of these measures, as these limit the achievable correlations across measures, whatever their interpretation. The data are also used to illustrate the problems of interpretation of RT and RT difference measures raised by Miller and Ulrich's (2013) individual differences in RT (IDRT) model.

* Corresponding author.
E-mail address: kenp@sfsu.edu (K.R. Paap).

http://dx.doi.org/10.1016/j.jneumeth.2016.10.002
0165-0270/© 2016 Elsevier B.V. All rights reserved.

1.2. Components of executive functions

Executive functions consist of a set of general-purpose control processes believed to be central to the self-regulation of thoughts and behaviors that are instrumental to accomplishing goals. The influential work of Miyake and Friedman (2012) used confirmatory factor analyses (CFA) to validate three of the components: updating, shifting (switching), and inhibition based on nine observed measures. At the higher level the three latent variables correlated with one another and this is consistent with an interpretation that each contributes to a common EF. When the same data were reanalyzed with a second order CFA such that the three latent variables were nested under a common EF latent variable, the nine observed measures all loaded on common EF with two of the components (updating and shifting) still making unique contributions. These findings support the assumption of a general EF ability with separable updating and switching components and an inhibition component that is not separable and that is weakly to moderately linked to the general EF ability. Because the best models of the data include both common and componential levels Miyake and Friedman propose that EF has both *unity* and *diversity*.

1.3. Convergent validity in measures of EF

This componential framework allows for the possibility that the related components have some degree of neural and functional independence. Thus, individuals may vary in terms of overall EF ability or with respect to specific components. If EFs are general-purpose then individuals who excel in, say, a measure of interference control in one task should also show little interference in a different task. That is, indices obtained in different tasks, but assumed to measure the same component of EF, should correlate and show convergent validity.

In order to isolate a component of interest, such as interference control, researchers often use the difference score between a condition that involves conflict (incongruent trials) and one that does not (congruent trials). For example, studies testing for bilingual advantages in interference control have usually relied on one or more of these interference tasks: Stroop color-word, Simon, spatial-Stroop, and flanker task. As discussed later (Section 4.4) in the context of the IDRT model (Miller and Ulrich, 2013) the subtraction does not purely isolate the process of interest (individual differences in the ability to resolve conflict), but here the concern is on the factors that limit and determine the test-retest reliability of these difference-score measures and the implications this has for assessing the convergent validity of two measures designed to reflect the same process of interest.

Turning to the reports of convergent validity, the intertask correlations obtained by Paap and Greenberg (2013) and Paap and Sawi (2014) and in the studies they review are disappointing at best. For example, Paap and Greenberg's Study 3 tested 107 participants and the cross-task correlation between the flanker and Simon effect was $r = -0.01$. Another striking example is that the two most popular versions of the flanker task, the letter version introduced by Eriksen and Eriksen (1974) and the arrow version first used by Stoffels and van der Molen (1988) do not correlate with one another as demonstrated by Salthouse (2010). In a study using 265 participants and 50 trials per condition, the correlation between the two flanker effects was +0.03. Low levels of convergent validity may be caused by one or both tasks having low reliability. Therefore, we consider next the determinants of reliability.

1.4. Determinants of reliability

In classical test theory the reliability of a measure derived from a mean is the correlation across individuals between two "parallel" measures, Corr[X, X′]. The measures are parallel in the sense that, for each individual, they assess the same true (T) score and have the same observed-score variance. Thus, the true-score variance between subjects combines with the average random error to determine reliability (Lord and Novick, 1968):

$$\rho_{XX'} = Corr\left[X, X'\right] = \frac{Var\,[T]}{Var\,[X]} \tag{1}$$

More specifically, reliability increases when there is more variability in the true scores between individuals and when there is less error variance. Thus, the reliability of a measure based on a single mean can be compromised by a lack of variability across subjects; for example, because the task is too easy or because the sample is very homogeneous. Alternatively, reliability can be compromised by increases in trial-to-trial variation caused by fluctuations in, for example, the participant's arousal, alertness, or motivation.

1.5. Effects of reliability on correlations between two measures

For any two simple measures X and Y, the correlation between X and Y must always be smaller than the square root of the reliability of the individual measures (Lord and Novick, 1968):

$$\rho_{XY} \leq \sqrt{\rho_{XX'}} \tag{2}$$

For example, if the test-retest reliability of measure X is 0.64, then the correlation between X and Y cannot exceed the upper bound of 0.80. Thus, the correlation between two measures can be compromised by a lack of reliability in either measure.

1.6. Reliability of several measures commonly assumed to index EF

Measures often assumed to index EFs appear to have adequate reliability, but the reported reliability is usually based on correlations between segments of a single session. For example, Paap and Greenberg (2013) reported block-to-block correlations and Spearman-Brown Prophesy[1] (SBP) values for ten assumed measures of EF. Eight yielded SBPs greater than 0.90. The two exceptions were SBPs of 0.734 for switching costs and 0.448 for the Simon effect. Although the correlation for the Simon effect is highly significant it is based on a very large sample size. Whether the correlation between segments of a single session is large or moderate, the primary purpose of this study is to determine if the test-retest reliability of these measures are systematically smaller when repeated on another day. This is plausible because performance on choice RT tasks is likely to be influenced by changes in arousal, attention, and motivation that are far more likely to vary across days than within testing sessions typically lasting less than an hour. Test-retest reliability across days or even longer spans is important if a measure is used to evaluate an individual's specific cognitive ability at a specific point in time or to infer the benefits of an intervention intended to hone the ability.

A comprehensive study reported by Wöstmann et al. (2013) provides apparently promising results on the test-retest reliability of several measures assumed to reflect inhibitory control: stop-signal, go/no-go, antisaccade, Simon, Eriksen flanker, Stroop, and Continuous Performance (CPT). With respect to measures of mean RT in a single condition the correlations range from $r = 0.55$ on the CPT to $r = 0.97$ for antisaccade latency. For difference-score measures (e.g., incongruent trial RT – congruent trial RT) like the flanker and

[1] SBP values are based on the correlation between two or more subsets of trials from a single session. The formula predicts the reliability that should be obtained if the number of trials is doubled, or more generally increased by a factor of n: $SBP = (n \times r)/(1 + (n - 1)r)$, where r is usually the Pearson correlation coefficient.

Table 1
Test-retest reliability for 18 RT measures derived from four different tasks across two sessions separated by about one week.

Task Measure	T/C	M 1	SD 1	M 2	SD 2	Test-Retest	p
Antisaccade							
Antisaccade RT	60	595	229	522	181	0.882	<0.001
Baseline RT	15	551	196	510	149	0.735	<0.001
Anti. RT Costs	60/15	44	101	11	88	0.442	<0.001
Flanker							
Neutral Baseline	20	469	68	444	61	0.642	<0.001
Incongruent Trials	96	574	58	541	51	0.879	<0.001
Congruent Trials	96	490	54	464	47	0.858	<0.001
Inc – Con	96/96	84	24	77	18	0.515	<0.001
Global RT	192	529	55	500	48	0.856	<0.001
Simon							
Neutral Baseline	20	457	47	442	41	0.654	<0.001
Incongruent Trials	40	482	46	464	44	0.719	<0.001
Congruent Trials	40	450	41	434	38	0.714	<0.001
Inc – Con	40/40	31	24	30	20	0.428	<0.001
Global RT	80	466	42	449	40	0.738	<0.001
Switching							
Single Task Baseline	32	550	211	523	138	0.773	<0.001
Repeat Trials	48	768	282	656	217	0.865	<0.001
Switch Trials	48	970	354	810	298	0.855	<0.001
Switching Cost	48/48	201	116	154	115	0.615	<0.001
Mixing Cost	48/30	218	258	133	185	0.745	<0.001

Note. T/C = trials per condition, M 1 = mean day 1, M 2 = mean day 2.

Simon interference effects the correlations were r = 0.94 and 0.71, respectively.

1.7. Selection of measures for present study

The present study involves participants completing the four tasks described in Section 2 and then returning, about a week later, to repeat the tasks again. Each task enables the derivation of both single-mean RTs and difference-score RTs that have been used to make inferences about the underlying processing. The purpose of this section is to explain why those tasks and measures were selected.

Additional study of the test-retest reliability of commonly used measures of interference control are needed because small changes in the task can alter the underlying processing and this, in turn, may also lead to differences in reliability. A good example is that the Simon task used by Wöstmann et al. is often referred to as a spatial Stroop task because in the Kornblum (1994) taxonomy it is usually classified as having both Stimulus-Stimulus (S-S) and Stimulus-Response (S-R) conflict (Egner, 2008), whereas our Simon task is usually classified as having only S-R conflict.

Similarly, as described in detail below our antisaccade task measures the latency of target identification when preceded by an opposite hemifield distractor whereas Wöstmann et al. measured the latency of a saccadic eye movement when preceded by an opposite hemifield distractor. Although both tasks may require the inhibition of a prepotent but counterproductive saccade, our task requires more task-specific central processing that could result in different test-retest reliabilities. Our task is based on one developed initially by Kane et al. (2001) and is frequently interpreted as a measure of inhibitory control (e.g., Unsworth and Spillers, 2010; Unsworth et al., 2012).

Our flanker task and the one used by Wöstmann et al. are both variants of the arrow-flanker task introduced by Stoffels and van der Molen (1988). The test-retest reliability of the flanker effect (0.94) reported by Wöstmann et al. is extraordinarily high for a difference-score measure and simply merits replication.

In addition to investigating the reliability of measures of interference control, we are also interested in the reliability of measures of the shifting component of EF. Many different tasks are used to assess general switching ability. The randomly-cued color-shape switching task (described in Section 2.6) was selected because of our long-standing interest in investigating the hypothesis that bilinguals are better general task switchers than monolinguals because of their ubiquitous practice in switching between languages (e.g., Paap, 2014). Our task was modeled on the seminal paper by Prior and MacWhinney (2010). As enumerated in Table 1 of Paap et al. (in press) 25 subsequent studies used the same or very similar instantiations of the task. The Prior and MacWhinney task, in turn, was modeled from the switching task developed in a highly cited paper by Rubin and Meiran (2005). Despite its ubiquitous use the present study is the only one reporting test-retest reliability across separate days. The most common measure of the shifting component of EF is referred to as switch cost and is defined as the mean RT on switch trials minus the mean RT on repeat trials within blocks requiring frequent task switches.

Another measure typically derived from the color-shape switching task is mixing cost defined as the mean RT on the repeat trials from a block that randomly mixes the two tasks minus the mean RT from single-task blocks. During the mixed block participants must prepare for a possible switch by maintaining both task sets (the stimulus-response rules) in working memory and must prepare for and then identify the precue even on the repeat trials when they do not switch tasks. Thus mixing costs reflect all processes that are required on both repeat trials and switch trials but that are not required when performing only a single task.

The present study will also focus on the reliability of global RT (the mean across both congruent and incongruent trials) in the flanker and Simon tasks as this measure is often used to investigate individual and group differences in the performance of nonverbal interference tasks. As discussed in Section 4.1 the IDRT model assumes that any single mean RT will be influenced by perceptual and motor differences as well as the central processing differences of interest, but in studies of interference control these concerns are sometimes ameliorated with experimental designs showing that the groups do not differ in a control block where there is never any conflict. A prominent example of the use of global RT is Hilchey and Klein's (2011) detailed review of dozens of studies using nonverbal interference tasks that showed no systematic differences between bilinguals and monolinguals in the magnitude of the interference effect, but a consistent bilingual advantage on both congruent and incongruent trials. This pattern of behavioural results together with the relevant neuroscientific data led Hilchey & Klein to propose that managing two languages leads to a system, based on early conflict monitoring, that can distribute inputs to separate processing centers for conflict versus non-conflict processing. This division of labor between functionally distinct processing streams and the consequent freeing up of process resources was assumed to be responsible for the ubiquitous global RT advantage.

In striking contrast, when Hilchey et al. (2015) updated their review they reported that the last two years of research dramatically challenged the evidence for bilingual advantages in global RT. In spite of this shift in the meta-analytic landscape and the task-impurity problem many researchers have claimed that global RT provides insights into bilingual advantages in monitoring. For example, Costa et al. (2008, p. 66) offer the following conjecture. *“Given that the task includes congruent and incongruent trials, participants have to evaluate the behavioural adjustments needed for each trial to determine the subsequently appropriate action. This continuous monitoring process in charge of detecting potentially conflicting information depends also on the executive control network, and consequently bilinguals should be overall better (faster) than monolinguals”* p. 66. In summary, global RT in combination with other RT measures is often used to make inferences about monitoring and provides a concrete example of the trade-offs between higher reliability (lower processing purity) in single mean RT measures and lower reliability (greater purity) in difference-score measures.

1.8. Source of data for present analysis of test-retest reliability

The test-retest reliabilities of the assumed EF measures reported below are derived from four different tasks that were completed in a first session and then repeated approximately one week later. The Session 1 results were reported by Paap and Sawi (2014) in a study that focused on testing for bilingual advantages in EF and on the convergent validity of measures commonly assumed to reflect the same component of EF.[2] No systematic differences between bilinguals and monolinguals were observed and both groups were combined in the present study in order to explore the test-retest reliability of each measure.

There are several research questions. To what degree and consistency are the test-retest reliabilities of frequently used measures of EF smaller than the reliability between segments from a single session? Further, to what degree and for what reasons are difference-score reliabilities smaller than those based on mean RTs in a single condition?

2. Methods

2.1. Participants

Paap and Sawi (2014) tested a cohort of 120 SFSU students who participated in order to fulfill a class requirement or for extra credit. The vast majority were upper division psychology majors. Eighty-one of the participants completed both sessions and the remainder chose to participate in only one session. The analyses reported below are restricted to the participants who completed both sessions.

2.2. Overall design and general procedures

During Session 1 each participant completed an informed consent form, the background questionnaire, and the four computer-controlled tasks in this order: antisaccade, flanker, Simon, and color-shape switching. If the participant elected to participate in two sessions, they returned approximately one week later and repeated the same four tasks in the same order.

2.3. The antisaccade (or distractor-suppression) task

The design, materials, and procedure for the antisaccade task was closely modeled on those used by Kane et al. (2001) who showed that individual differences in working-memory capacity predicted performance on an antisaccade blocks, but not prosaccade blocks. The task on each trial was to identify the target stimulus (i.e., “B”, “P”, or “R”) by pressing the key (from a side-by-side row of three) with the corresponding label using three fingers of the right hand. The briefly presented target is followed by a visually similar mask (“8”). The target and mask subtended about 0.9 ° of visual angle. In the antisaccade condition a distracter stimulus is always blinked (presented for 100 ms with a blank ISI of 50 ms) before, and on the opposite side from, the target stimulus. The distractor appeared about 2.0 ° to one side of fixation and the target 2.0 ° to the opposite side. Because the eventual target is always presented on the opposite side the best strategy is to inhibit the natural predisposition to attend to (or saccade toward) any peripheral stimulus with an abrupt onset. Individuals with superior inhibitory control, should respond faster in the antisaccade task. The antisaccade trials are preceded by a block of control trials that used a centered target and no distracting stimulus. The control trials provide a baseline response time (RT) that should require no inhibitory control or distractor suppression.

Experimental trials consisted of the following sequence of events: (1) a center fixation (***) was presented for a variable duration (i.e., 600, 1000, 1400, 1820, 2200 ms) in order to introduce temporal uncertainty; (2) a blank field for 100 ms; (3) a “#” sign for 100 ms displaced 2 ° to the opposite side from the eventual target; (4) a blank field for 50 ms; (5) the “#” sign in the same location for 100 ms; (6) a target letter (“B”, “P”, or “R”) for 150 ms displaced a comparable extent on the opposite side; (7) a mask (‘8’) presented until the response.

The baseline trials presented no opposite field distracter and consisted of these events: (1) a center fixation (***) was presented for a variable duration (i.e., 600, 1000, 1400, 1820, 220 ms; (2) a blank field for 100 ms; (3) a centered target-letter (“B”, “P”, or “R”) for 150 ms; and (4) a mask (‘8’) presented until the response.[3]

The trials were organized and presented in the following order. A practice block consisted of 15 baseline trials, one at each combination of 5 fixation durations and 3 target letters and presented in random order. Block 2 was identical to the first block and provided the baseline RTs. Block 3 was 30 antisaccade trials formed by the random combination of: 5 fixation durations by 3 target letters by 2 sides (left and right).

2.4. The ANT task

A common measure of interference control is the flanker effect as described in the introduction. However, an extension of the basic

[2] The Paap and Sawi (2014) was an invited paper, with a strict page limit, on the topic of the alleged bilingual advantage in EF and the format did not allow room for inclusion of the second session data and the issue of test-retest reliability.

[3] We used this centered condition rather than a prosaccade condition, and always tested it first, because Kane et al. showed that low-span participants performed poorly when switching from antisaccade to prosaccade blocks. Our goal was to establish a neutral baseline for identifying masked targets in the absence of exogenous distractors.

flanker task developed by Fan et al. (2002) is also popular and generates measures that are claimed to reflect alerting and orienting, and inhibitory control. The extended task is referred to as the attentional network task (ANT). The particular version of the ANT task used by Paap and Sawi (2014) and in the present assessment of test-retest reliability replicates the design and procedures used by Costa et al. (2008).

The cues, arrows, and flankers were implemented in DirectRT using Costa et al.'s (2008) Figure 1 as the model. The congruent display consisted of a central arrow pointing either left or right and two flankers on each side pointing in the same direction as the central target. A single arrow subtended about 0.9 ° of visual angle and the entire horizontal extent of the five-arrow stimulus was about 6.3°. In the incongruent displays the flankers pointed in the opposite direction from the central target arrow. The sequence of events was as follows: (a) a fixation point (a plus sign) appeared at the center of the screen and remained throughout the trial, (b) a cue (described below) was presented for 100 ms, (c) followed by the fixation field for an additional 400 ms, and then (d) the target display until the participant's response or for up to 1700 ms. The target was vertically displaced either 1.2 ° above or below the fixation point. Participants were instructed to press the "z" key with their left index finger if the target arrow pointed left and to press the "/" key with their right index finger if the target arrow pointed right.

Consistent with the ANT methodology four types of cues were used. On "no cue" trials the 100 ms cue display is simply a continuation of the centered fixation point (+). Obviously it affords no information about the temporal onset or spatial location of the upcoming target. The "double cue" display consists of a two ◊ symbols above and below the fixation point. This provides no information about the location of the upcoming target, but does reduce the temporal uncertainty. Subtracting the means of the double cue trials from the no cue trials yields the alerting effect. The third type of cue is the "central cue" that simply replaces the + fixation point with the ◊ symbol. It does reduce temporal uncertainty, but provides no cue to spatial location. In contrast, the "spatial cue" display adds a valid diamond cue above or below the fixation point. As both the "central cue" and "spatial cue" displays provide the same advantages in alerting, the mean of the "spatial cue" trials can be subtracted from the mean of "central cue" trials to derive the orienting effect.

Block 1 consisted of 20 neutral trials where all the targets consisted of a centered arrow and the flankers were dashes. Each target was randomly preceded by one of the four cue types. Blocks 2 through 5 were standard ANT blocks with 50% congruent and incongruent trials. Block 2 consisted of 16 trials and was considered practice. Blocks 3, 4, and 5 each consisted of 64 trials with 8 repetitions of the combinations formed by 2 target types (congruent versus incongruent) × 4 cue displays. Thus, given standard practice for analyzing each attentional network (executive attention, alerting, and orientating) in the ANT each block provided 32 trials of each condition (e.g., 32 congruent and 32 incongruent trials) and overall means collapsed across blocks were based on 96 trials. The trials within each block were randomized.

2.5. The Simon task

The Simon task was identical to one used by Paap and Greenberg (2013) in their Studies 2 and 3. Each trial began with the presentation of a center fixation (+) for 500 ms. The center fixation was immediately followed by the target stimulus which was either a "Z" or a "/". The participant's task was to press the corresponding key on the computer keyboard as quickly as possible without making errors. The left index finger rested on the "Z" key and the right index finger rested on the "/" key. In the Simon blocks the target was displayed either 3.9 ° to the left or to the right of the center fixation. In the Simon blocks a trial was defined as congruent if the location of the target was on the same side as the correct response and as incongruent if the location of the target was on the opposite side.

The critical Simon blocks were always the last two of four blocks. Each Simon block consisted of 20 congruent trials and 20 incongruent trials presented in random order. Half the trials of each type presented the target on the left with the other half presented the target on the right. Thus, the mean response time (RT) for the four conditions defined by the combination of two blocks and two levels of congruency (congruent versus incongruent) were each based on 20 trials and when collapsed across blocks on 40 trials.

Prior to the Simon-task blocks there were two blocks of trials where the target was displayed either 2.3 ° above or below the center fixation. This change introduces a "neutral" condition because the location of the target is neither compatible nor incompatible with pressing the "Z" key on the left or the "/" key on the right. Block 1 provided 20 trials of practice in the neutral condition and was followed by a 40-trial Block 2. Displacements above and below the fixation were randomly ordered with the constraint that within the block there were 20 displacements above and 20 below. The Simon-task blocks followed the neutral blocks.

As the neutral (no conflict) control condition always preceded the mixed block the difference between the neutral RT and the congruent RT from the mixed block may underestimate the cost of mixing (mean congruent RT from mixed block minus mean RT in the control block) in both the flanker and Simon task. This potential problem with practice effects may have been attenuated by using a sandwich design in which pure- and mixed-blocks are alternated.

2.6. The color-shape switching task

The task was patterned on that used by Prior and MacWhinney (2010) in their seminal investigation of the relationship between bilingualism and task-switching ability. In turn, Prior and MacWhinney based their task on a paradigm developed by Rubin and Meiran (2005). Each trial began with the presentation of a center fixation (+) for 350 ms and then a blank screen for 150 ms. The left middle and index fingers rested on the "Z" and "X" key, respectively. The right index and middle fingers rested on the "." and "/" keys, respectively. In a pure color block the participant's task was to press the "Z" key if the target was blue and the "X" key if it was red. In a pure shape block the task was to press the "." key if the target was a circle and the "/" key if it was a triangle. The target set consisted of a blue circle, a blue triangle, a red circle, and a red triangle.

In a mixed block the target was preceded by a precue for 250 ms that remained in view until the participant responded to the target. If the precue was a rainbow, then the participant had to make a color decision when the target appeared. If the precue was a black circle embedded within a black triangle, then the participant had to make a shape decision when the target appeared. Participants were instructed to respond as quickly as they could on the basis of the precued dimension n (viz., color or shape). Each trial was designated as a "repeat" trial if the cued dimension was the same as on the previous trial and a "switch" trial if it was different. Each target and precue subtended about 1.83 ° of visual angle with the center of the precue appearing 2.3 ° above the center of the fixation stimulus and the upcoming target.

The task consisted of six blocks. The first block of 16 trials was "pure" color. Each of the four targets appeared four times in random order. The second block of 16 trials was "pure" shape with each of the targets appearing in random order. Following Block 2 the "mixed" task was introduced with detailed instructions regarding the use of the precue to signal whether a color or shape would be required on each specific trial. Each of the four "mixed" blocks started with two buffer trials that were not analyzed. Block 3 was a practice block and consisted of 18 trials (including the two buffers). Blocks 4, 5, and 6 each consisted of 50 trials (including the two buffers). A single random order was used for every participant. Each of the four targets appeared 36 times across Blocks 4 to 6 and there were 72 repeat trials and 72 switch trials.

It bears mentioning that switch costs in the task used here do not provide a pure measure of task-set reconfiguration because the task cue appears on every trial at a relatively short SOA and there is only a single cue associated with each task (Monsell and Mizon, 2006). Although this switching task may not be the best tool for measuring task reconfiguration it has the virtue of providing a context and imperfect standard for dozens of published studies testing for individual or group differences in switch costs.

3. Results and discussion

The main goal of this study is to assess the test-retest reliability of RT measures often associated with EF and the restrictions that the reliabilities have on the convergent validity of measures assumed to reflect the same component of EF. Participants were quite accurate in all four tasks, the mean proportions correct were 0.98, 0.97, 0.94, and 0.93 for the flanker, Simon, switching, and antisaccade tasks, respectively. Consistent with this goal we focus on the RT measures.

For the RT analyses the standard deviation (SD) for the experimental trials of each individual participant were computed and RTs that exceeded 2.5 SDs were trimmed. The data from five participants were removed from the analyses because their accuracy levels on the antisaccade task were near chance levels (viz., <0.5 with three alternatives). The data from one participant was removed because this individual showed a large negative Simon effect (−54 ms) in Session 1, a value 3.5 standard deviations less than the mean of 32 ms. To maintain population equivalence across the tasks the data from these six participants were also removed from the analyses of the other three tasks. Thus, the test-retest reliabilities are based on 75 participants for all tasks.

The means, standard deviations, and test-retest reliabilities of 18 measures are shown in Table 1. The difference-score measures are shown in italics. The number of trials per condition (T/C) is also shown. If the measure is a difference score, then the T/C is shown for both the more difficult and less difficult condition. As discussed in detail in Section 4.4.2 on the IDRT model, reliability increases with T/C because it reduces the error variance.

As a general observation of the results shown in Table 1, correlations between measures that consist of a single RT have a median of 0.77 ($r^2 = 0.60$), whereas correlations between differences scores have a median of 0.51 ($r^2 = 0.27$). High levels of statistical significance are not very helpful in evaluating the score obtained by individual participants. Given the relatively high sample size in this study a correlation of 0.3 is significant with $p < 0.01$; but only 9% of the variation in a second round of testing can be accounted for on the basis of the knowledge of scores in the first round. Small, but significant levels of test-retest reliability do enable one to test hypotheses concerning group differences but a low (or even moderate) level of test-retest reliability signals the need for larger sample sizes in order to achieve adequate power and minimize Type 1 and Type 2 errors.[4] The rate of false positives is inflated when a field routinely uses small n's (Bakker et al., 2012; Button et al., 2013; K.R. Paap et al., 2014; Paap et al., 2015).

3.1. Is target-identification latency in our antisaccade task a reliable measure?

Our antisaccade task required participants to ignore a briefly flashed distractor and identify a target presented to the opposite side. The mean RT in a pure block of antisaccade trials was used as a measure of inhibitory control. As shown in Table 1 the test-retest reliability of the *antisaccade RT* measure was r = 0.88. Because the design included a block of baseline trials where there was no distractor and the targets were presented at fixation, a second measure of inhibitory control subtracts the mean RT on the block of baseline trials from the mean RT on the block of antisaccade trials. The measure is referred to as *antisaccade costs* and its test-retest reliability was r = 0.44. Although highly significant, the low value compromises its potential to establish convergent validity with other measures of inhibitory control imposing a maximum cross-task correlation of 0.65.

There are no other reports of test-retest reliability for antisaccade tasks that are based on the speed of target identification, but our results can be compared to studies measuring the latency of saccadic eye movements opposite a visually presented "go" stimulus. For example, our *antisaccade RT* measure compares well with the test-retest reliability of r=0.91 reported by Wöstmann et al. and is higher than the reliability of r=0.77 reported by Klein and Fischer (2005). Although these investigations also ran prosaccade conditions, there were no neutral conditions that would enable a measure of antisaccade cost comparable to ours.

3.2. Is the Simon effect a reliable measure?

The reliability of the standard Simon effect (r = 0.43) was significant but small and may be inadequate for many purposes. Only 18% of the variance in Day 2 Simon effects can be accounted for on the basis of the observed Day 1 effects. Also, the reliability of 0.43 limits the degree of convergent validity (see Eq. (2)) between the Simon effect and other measures of interference control to 0.66.

The results reported by Wöstmann et al. (2013) for their "Simon task" are very different. Their Day 1 Simon effect (M = 71, SD = 53) is about twice as large (M = 32, SD = 41) with a much higher test-retest reliability, 0.71 compared to 0.43 in our results. The larger interference effect is expected for two reasons. First, as noted earlier the Wöstmann et al. task can be classified as a spatial Stroop task that, unlike the standard Simon task, includes S-S interference. Second, there are twice as many congruent trials (160 per session) compared to incongruent trials (80 per session) and interference effects tend to increase when incongruent trials are less expected. To the extent that these task differences increase task difficulty and enable more true score variability, the greater test-retest reliability reported by Wöstmann et al. makes sense. A more general conclusion is that the Simon (used by us) and spatial Stroop (used by Wöstmann et al.) should probably be considered as two different tasks that depending on how each is instantiated are likely to have different test-retest reliabilities. To the extent that they resolve conflict at different levels of processing, they are also likely to show different patterns of convergent validity with other measures of interference control.

3.3. Is the flanker effect a reliable measure?

The reliability of our flanker effect (r = 0.52) was somewhat higher than for our Simon effect. However, it is far smaller than the surprisingly high test-retest reliability reported by Wöstmann et al. (2013), r = +0.94. As developed later in the discussion we consider the value reported by Wöstmann suspect. However, there are substantial differences in how the flanker task was instantiated and

[4] Larger samples sizes are, of course, not the solution to low reliability if the root problem is population homogeneity.

in the magnitude of the flanker effect. We used a fairly standard version of the ANT task with arrows that look similar to this, →, whereas Wöstmann et al. used a standard (no alerting or orienting precues) flanker with targets and flankers that looked more like this, ▷. Also, our neutral trials were presented only in a block of baseline trials where conflict never occurs rather than being mixed into the experimental block. The net result of these differences in task and population was that our Day 1 flanker effect (M = 83, SD = 26) was greater than that of Wöstman et al.'s (M = 32, SD = 32). If our larger flanker effect enabled greater true score variation, then this factor would predict greater reliability. Thus, consideration of relative magnitude of the flanker effects only deepens the puzzling discrepancy.

Ishigami and Klein (2010) reported test-retest reliabilities for flanker effects (called executive control in the ANT literature) derived from the original ANT (Fan et al., 2002) and a version developed by Callejas et al. (2005) that used a tone cue rather a visual cue to enable the temporal alerting condition. There were 10 young adult participants and the average test-retest interval was 8.6 days. The reliability for the original ANT was large and significant ($r = 0.86$), but smaller and non-significant for the tone-cue ANT ($r = 0.48$). Given that the target displays for congruent and incongruent trials were identical in the two versions, the discrepancy is somewhat disconcerting.

In Ishigami and Klein (2011) a group of older adults (M = 69 years) were retested across an average interval of 6.7 days and the reliability was 0.57 (nonsignificant) for the standard ANT and 0.79 ($p < 0.01$) for the tone-cue version. In both studies and both versions of the ANT task there were 96 trials per condition. The fact that the test-retest reliability of the flanker effect was greater for standard ANT with young adults and was greater for the tone-cue ANT with the older adults probably reflects volatility caused by small samples (Paap et al., 2015).

Although the focus of this article is on test-retest reliability across retest intervals of days or weeks, it is worthwhile to note that MacLeod et al. (2010) reviewed 15 unique studies using the ANT (resulting in a large sample of 1129 individuals) and reported an average split-half reliability for the flanker RT effect of $r_{\text{weighted}} = 0.65$, CI $95\%_{\text{weighted}}[0.61, 0.71]$.

3.4. Is switching cost a reliable measure?

The differences between the repeat trials and switch trials from the block where the required decision is precued during each trial are referred to as *switching costs* and are usually assumed to reflect the efficacy of the ability to switch. As shown in Table 1 the test-retest reliability for *switching costs* is $r = 0.62$; a relatively high level of reliability for a difference-score measure.

The reliability of switching costs derived from the color-shape task is similar to that obtained from other tasks assumed to measure switching ability. This is noteworthy given that the examples provided below are for measures based on the mean of a single condition rather than a difference score. For example, the D-KEFS Trails task that requires participants to alternate between connecting digits and letters yields a reliability of $r = 0.59$ over an average test-retest interval of 25 days (Delis et al., 2001). A similar reliability ($r = 0.65$) was observed for the D-KEFS color-word switching condition in which participants must alternate between naming the color and reading the word for a set of incongruent Stroop materials. The Shifting Attention Test (SAT) is part of a computerized battery of EF tasks (Gualtieri and Johnson, 2006). The SAT measures the ability to shift from one instructional set (match by shape) to another (match by color) quickly and accurately. The rules change at random and in that regard the task is similar to the Wisconsin Card Sort. The test-retest reliability of the SAT, across a median interval of 27 days was 0.71.

3.5. Is mixing cost a reliable measure?

As noted in the introduction researchers using the color-shape (or similar) switching task frequently compute "mixing cost" that are computed as the difference between the mean of the single task (pure color or pure shape) trials and the repeat trials from the mixed block. The mixing cost measure showed a test-retest reliability of 0.74 which is impressive for a difference-score measure.

3.6. Is global RT a reliable measure?

Global RT is highly reliable in both our Simon task ($r = 0.74$) and flanker task ($r = 0.86$), but using global RT as a measure of monitoring is risky because all task processes influence global RT. The task impurity problem is discussed in the next section.

4. An IDRT analysis of the tradeoffs between RT differences scores and RTs from a single condition

A well-known problem in interpretation is that no single task embodies any single process such as inhibitory control (Burgess, 1997; Paap, 2014). One common strategy for dealing with task impurity is to form a difference score that ideally isolates the process of interest (e.g., inhibitory control) by subtracting out all the shared processing components involved in such things as perceptual encoding and response execution. The tradeoff between the greater purity of difference-score measures and the greater reliability of single-mean measures has been clarified in an elegant modeling and parameter-exploration project by Miller and Ulrich (2013). In this section the model is described and applied to the interpretation of the measures reported earlier.

4.1. The core assumptions of the IDRT model

The component processes of generic choice-RT tasks and their associated individual differences are captured by Miller and Ulrich in their IDRT model. The observed mean reaction time RT_k of a single participant k is the sum of the latencies of several processing stages:

$$RT_k = (A + B + C) \cdot G_k + B \cdot \Delta_k + R_k + E. \quad (3)$$

The constants A, B, and C represent the workload (dependent on the task not the person performing it) associated with perceptual input (constant A), central processing (constant B), and motor output stages (constant C), respectively. The weight on this sum, G_k, represents individual differences in general *processing time* (likely related to overall neural processing speed). The usual focus of theoretical interest is Δ_k, the processing time parameter reflecting participant k's particular ability at the central processing (component "B") required in a particular experimental condition. It is worth noting that Δ_k is likely to be different from k's general processing speed, G_k, and reflects k's specific facility to carry out B (e.g., k's ability to select the correct response in the presence of conflict between the flankers and target on an incongruent flanker trial). R_k represents individual variation in perceptual and motor processing that is constant across conditions such as the time required for light transduction processes within the retina. The last component, E_k, is a statistical error term caused by random trial-to-trial variability of RT for a given participant in a given condition. The error variance, $Var[E_k]$, differs across individuals and can be estimated for each individual as the standard error of the mean for that participant which, in turn, depends on the variability of that participant's

RT distribution and on the number of trials for each participant. Var[E] can be estimated by averaging the individual standard errors in a condition. For example, the estimate is 7.9 ms for global RT in our Simon task compared to 6.3 ms for global RT in our flanker task. The greater error variance in the Simon task occurred because this task had only 80 trials compared to 192 for the flanker task as the individual standard deviations were actually greater in the flanker task.

4.2. Reliability of mean RTs in the IDRT

The generic expression for the reliability of any score X that was introduced in Eq. (1) can be rewritten for scores representing mean RTs:

$$\rho_{RT,RT'} \equiv Corr\left[RT, RT'\right] = \frac{Var\,[T]}{Var\,[RT]} = 1 - \frac{Var\,[E]}{Var\,[RT]} \tag{4}$$

Eq. (4) implies that reliability increases with increases in the variance of the true scores (Var[T]) and with decreases in the error variance (Var[E]). To take a concrete example the somewhat lower test-retest reliability for global RT in our Simon task (r=0.74) compared to our flanker task (r=0.86) is mostly due to the smaller number of trials increasing the error variance as discussed in the preceding paragraph.

When comparing test-retest reliability across tasks performed by the same participants, the variance of the true scores depends only on the differences between the tasks. In general increasing task difficulty will increase true score variability, but it also increases the error variance resulting in a tradeoff. That error variance increases with task difficulty follows from the observation that mean RT and mean SD usually correlate. For example, these correlations are 0.73 and 0.67 for global RT in the flanker and Simon task, respectively. This relationship is often expressed as the coefficient of variation (CV):

$$CV = SD_{RT}/Mean_{RT} \tag{5}$$

The CVs for our flanker (CV = 0.16) and Simon tasks (CV = 0.15) are quite similar. Thus, the more difficult flanker task (15% longer mean RT) probably has somewhat higher error variance and somewhat greater true score variance. Interestingly, Miller and Ulrich (see their Figure 1) show that for a range of plausible parameter combinations that cover our RT measures the reliability of mean RT is influenced very little by either task difficulty (the values of parameters A and B) or by the CVs when the number of trials per condition exceeds 40. According to the IDRT the reliabilities should exceed 0.9. The fact that the test-retest reliabilities for global RT in the Simon and flanker task are “only” 0.66 and 0.73, respectively, suggests a possible violation of an assumption of the classic model, a possibility explored later.

The variance in true scores is a composite of the parameters G, Δ, and R and another important outcome of the IDRT explorations is that reliability increases with increases in any of the population variances: σ_G, σ_Δ, and σ_R. However, as Miller and Ulrich point out this is not always good because researchers wanting to test hypotheses about task-specific processing time (Δ) may be using a RT measure that enjoys high test-retest reliability only because of appreciable true differences in G and R. In summary, mean RTs will generally have high reliability with only a moderate number of trials, but in some cases the true score variance in the component of interest (e.g., task specific processing, Δ) may be insufficient to correlate highly with some other measure (e.g., the amount of training on the task).

4.3. Correlations between two mean RT measures

Consider a correlation between any two conditions assumed to require interference control (e.g., antisaccade RT, incongruent Stroop trials, incongruent flanker trials, or incongruent Simon trials). For example, the correlation between incongruent Simon trials (*SimonI*) and incongruent flanker trials (*FlankerI*) might be used to assess the extent to which these tasks either tap into different mechanisms for conflict resolution (as the Kornblum taxonomy predicts) or to validate their equivalence (as often assumed in meta-analyses of interference scores). The observed correlation is ambiguous because its value reflects not only the correlation between the underlying task-specific components ($\rho_{\Delta_{SimonI}\ \Delta_{FlankerI}}$), but also the underlying correlations in general processing time, ($\rho_{G_{SimonI}\ G_{FlankerI}}$) and in residual perceptual and motor processing time. General processing time is by definition common to both tasks and should drive a positive correlation that increases with the amount of true score variance in G. In contrast, the individual differences in central processing, Δ, are task specific and will be correlated if the processing is much the same in both tasks or uncorrelated if they are not.

Miller and Ulrich explored the space of possible correlations for a wide range of plausible parameter values. As shown in their Figure 4, the correlation between two mean RTs will always be high even when the true correlation on the task-specific parameter Δ is set very low ($\rho_\Delta = 0.075$). The overestimation increases with the true variance in general processing time. As Miller and Ulrich conclude “*. . .the correlation between the two RTs is not a good index of the correlation between the two task-specific processing times. . .*” p. 826. In other words, a significant correlation between, say, *FlankerI* and *SimonI* (r = 0.43 in our data) cannot be taken as strong evidence that the correlation between the underlying task-specific components ($\rho_{\Delta_{SimonI}\ \Delta_{FlankerI}}$) is substantial as the observed correlation can easily be driven by the underlying correlations in general processing time ($\rho_{G_{SimonI}\ G_{FlankerI}}$). This possibility could be ruled out if the correlations between the congruent conditions were weak. However, the correlation between *FlankerC* and *SimonC* in our data is also strong (r = 0.46). Thus, it appears that the general processing component of each task may be sufficient to account for the significant correlation between the incongruent conditions of the two tasks.

Another informative comparison involves the correlations between the congruent and incongruent trials within each task (r = 0.86 for the Simon task and r = 0.91 for the flanker task). These very high correlations suggest that individual differences in general processing time together with individual differences in task-specific processing unrelated to conflict resolution are driving the correlation. The last point merits elaboration. Both of these nonverbal interference tasks require a central process that maintains an S-R rule and then selects the appropriate response given the identified stimulus. In our study the Simon rule involves the identification of a symbol (“Z” or “/”) and the selection of the correct response in the presence of either congruent or incongruent spatial location information. In contrast, the flanker rule involves the identification of the direction of a centered arrow in the presence of either congruent or incongruent flankers. The high correlations between the congruent and incongruent conditions within each task suggest that individual differences in these two tasks are not perturbed by the presence or absence of conflict processing. If RTs on incongruent trials were substantially influenced by underlying differences in inhibitory control that are different from the underlying differences on the processing components shared between the two trial types, then the correlations between the congruent and incongruent conditions should not be high. Given that they are high the significant correlations between mean *FlankerI* and *SimonI* seem to have little or nothing to do with individual differences in conflict processing. Mean RTs in incongruent conditions are likely to be very impure measures of individual differences in interference control.

4.4. The reliability and purity of an RT difference score

4.4.1. Representing difference scores in the IDRT

Taking the difference in mean RT between an incongruent condition where conflict is present (subscripted as *inc* in the equations below) and a congruent condition where it is absent (subscripted as *con*) theoretically disentangles task-specific central processing associated with S-R mapping from that associated with conflict resolution under the assumption that the S-R mapping is the same in both conditions. The difference score can remove some of the influence of general processing time and give more influence to task-specific processing, Δ_{inc}, but it does not completely isolate Δ_{inc} as shown in the following equations. The first step is to express the mean RT in a given condition as a random variable[5] (e.g., $\mathbf{RT}_{inc}$ for the incongruent condition) rather than as the mean RT_k for individual k. Thus, the observed mean difference score (**D**) is written:

$$\mathbf{D} = \mathbf{RT}_{\text{inc}} - \text{RT}_{\text{con}}. \tag{6}$$

The terms from Eq. (3) can be substituted into Eq. (6):

$$\mathbf{D} = [(\text{A}_{\text{inc}} + \text{B}_{\text{inc}} + \text{C}_{\text{inc}})\cdot\mathbf{G} + \text{B}_{\text{inc}} \cdot\Delta_{\text{inc}} + \mathbf{R}_{\text{inc}} + \text{E}_{\text{inc}}] - [(\text{A}_{\text{con}} + \text{B}_{\text{con}} + \text{C}_{\text{con}})\cdot\mathbf{G} + \tag{7}$$

$$\text{B}_{\text{con}}\cdot\Delta_{\text{con}} + \mathbf{R}_{\text{con}} + \text{E}_{\text{con}}]$$

Under the reasonable assumption that A, C, and **R** are the same in both the experimental and control conditions this reduces to:

$$\mathbf{D} = [(\text{B}_{\text{inc}} - \text{B}_{\text{con}})\cdot\mathbf{G}] + [\text{B}_{\text{inc}}\cdot\Delta_{\text{inc}} - \text{B}_{\text{con}}\cdot\Delta_{\text{con}}] + \text{E}_{\text{inc}} - \text{E}_{\text{con}}. \tag{8}$$

Eq. (8)[6] shows that the mean observed difference score does not isolate the task-specific processing, $\boldsymbol{\Delta}_{inc}$, unique to the incongruent condition (e.g., the conflict resolution mechanism on incongruent flanker trials). It will also be influenced by the task-specific processing associated with the control condition, $\boldsymbol{\Delta}_{con}$ and by general processing time differences, **G**.

Miller and Ulrich show that these additional influences are far from trivial and critically depend on the directional relationship between the incongruent and control conditions. In some cases, the task-specific processes are *opposing* in that they have opposing consequences. For example, on incongruent Stroop trials the automatic word-reading response competes with the color-naming response and slows processing (i.e., increases $\boldsymbol{\Delta}_{exp}$). In contrast, on the congruent trials the automatic processing of the task-irrelevant word tends to speed responses. Thus, within IDRT, this difference entails a strong negative correlation between $\boldsymbol{\Delta}_{con}$ and $\boldsymbol{\Delta}_{inc}$. Similar task analyses lead to the expectation that the interference effects (differences between congruent and incongruent trials) in the Simon, spatial-Stroop, and flanker tasks also involve opposing task-specific processing. In contrast, for antisaccade costs the task-specific processing in the neutral baseline is likely to be unrelated to that in antisaccade condition. However, Miller and Ulrich caution that there is no way to empirically establish this relationship (the value of $\rho_{\Delta_{inc}\ \Delta_{con}}$) and that theoretical task analyses can lead one astray. Having said that, there may be good clues. For example, if a flanker tasks includes a random mix of congruent, incongruent, and neutral trials and if there is no facilitation effect (RT congruent ≈ RT neutral), then one might consider the possibility that the redundant flankers are not being used to speed task-specific processing on the congruent condition and that the relationship between the task-specific components may be unrelated rather than opposing.

4.4.2. Reliability of difference scores

Consider a difference-score measure (D) such as the difference in mean RTs between incongruent (Inc) and congruent (Con) trials: D = Inc – Con. Miller and Ulrich show that reliabilities tend to be substantially higher for difference scores based on opposing processes than for those based on shared (or unrelated) processes. Under a wide range of parameter combinations, hundreds of trials per condition are sometimes needed to produce reliabilities exceeding 0.8 (see Ulrich and Miller's Figure 5 for details), but thousands of trials per condition may be needed when the two conditions involve shared task-specific processing (rather than being opposing), especially if the effect size (i.e., values of B_{inc} compared to B_{con}) is not very large. A silver lining to the need for more trials per condition to achieve high levels of reliability for an RT difference score is that a high test-retest correlation (if one is fortunate enough to obtain one) is less ambiguous than a high correlation for a simple RT measure. This is because a high reliability for a mean RT measure can be caused by large variability in the sensory-motor residual component alone, **R**; but that cannot happen for a difference RT because the R component is subtracted out as shown in Eq. (8).

The preceding analysis adds perspective to the test-retest reliabilities of the difference scores in our study. Take for example, the significant but rather small reliability for our Simon effect (r = 0.43) and flanker effect (r = 0.52) that were based on 40 and 96 trials per condition, respectively. Given the parameter explorations provided by Miller and Ulrich in their Figure 4, these test-retest reliabilities are quite low. This is particularly true under the assumption that the congruent and incongruent conditions are opposing in both tasks. Under this assumption reliability should be no less than 0.80 even when the trials per condition are as low as 40. Alternatively, it may be that the congruent trials are closer to neutral than to opposing. Under this assumption reliabilities as low as 0.70 might be expected. It appears that the magnitude of the observed reliabilities is limited by another factor in addition to the number of trials per condition. One culprit implicated in both our data and Wöstmann's is likely "day" effects and this problem is discussed next.

4.4.3. Day effects sabotage test-retest reliability

The reliability of a difference score, Corr[D, D′], is subject to constraints from both the reliability of the individual measures Con and Inc and the cross-correlation between them. Under plausible assumptions for RT research (see Miller & Ulrich) the reliability of a difference-score measure reduces to[7]:

$$\rho_{\text{D},\text{D}'} = \text{Corr[D, D']} = \frac{\rho_{\text{ConCon}'} - \rho_{\text{ConInc}}}{1 - \rho_{\text{ConInc}}} \tag{9}$$

A consequence of Eq. (9), not always appreciated, is that the reliability of a difference score is inversely related to the cross-correlation between Con and Inc. In the limit when the cross-condition correlation, ρ_{ConInc}, equals the same-measure reliability (e.g., $\rho_{\text{ConCon}})_{'}$, the reliability of the difference score will be zero. This relationship is important for many common measures of interference control because the congruent and incongruent trials are often highly correlated as we have already discussed for both our flanker and Simon tasks.

[5] Random variables are indicated by bold font.

[6] In order to reduce the formula for the reliability of difference scores to Eq. (8) it is also necessary to assume that the congruent and incongruent conditions have equal variances and that the two conditions have equal reliabilities.

[7] The expression shown in Equation also holds for both the reliability of congruent RTs (as shown) and for the reliability of incongruent RTs, ρ_{IncInc}.

For our flanker data the underlying correlation between the congruent and incongruent trials (ρ_{ConInc}) can be estimated (with sampling variability) as r = 0.93. The reliabilities of the condition means are estimated as r = 0.48 for the incongruent trials and r = 0.53 for the congruent trials or about 0.50 on average. Substituting these values in Eq. (9) yields:

$$\rho_{D,D'} = \mathrm{Corr}[D, D'] = \frac{\rho_{ConCon'} - \rho_{ConInc}}{1 - \rho_{ConInc}} = \frac{.50 - .93}{.07} < 0 << 0.48 \tag{10}$$

There is an even greater disconnect between the reliability of the flanker effect of r = +0.94 reported by Wöstmann et al. and Eq. (9). The reliability of each condition mean can be estimated (with sampling variability) as 0.79 for the congruent trials and 0.87 for the incongruent trials or about 0.84 on average. The correlation between the congruent and incongruent trials was not reported by Wöstmann et al. but can be estimated from the means and SDs they report for the two trial types by using Equation 37 from Miller and Ulrich. Doing so yields an estimate of ρ_{ConInc} = 0.89. Substituting these values into Eq. (9) yields the negative correlation shown in Eq. (11).

$$Corr(D, D') \approx \frac{.84 - .89}{1 - .89} < 0 << 0.94 \tag{11}$$

It is nonsensical that the test-retest reliability of the flanker effect should be negative. The problem, of course, is that the correlations between measures (congruent and incongruent trials) within a session are greater than the test-retest reliability of the individual measures (Con with Con′ and Inc with Inc′) across sessions. This suggests that a "day effect" is operating.[8] Changes in arousal, alertness, motivation, or learning consolidated between sessions might all be changing relationships that classical test theory assumes are stable. The crux of this possibility is that the value of ρ_{ConInc} has an advantage over the values of ρ_{ConCon} or ρ_{IncInc} because both measures involved in ρ_{ConInc} are accumulated on the same day and presumably under the same circumstances (e.g., same levels of alertness, arousal, motivation, and learning). Day effects are likely responsible for the negative estimated Corr (D,D′) values, but there is no existing extension of classical test theory that takes the day effect into account and computes what the test-retest reliability would be in their absence. Furthermore, as shown in Figure 14 of Miller and Ulrich when true values of ρ_{ConInc} and ρ_{ConCon} both vary in proximity to 0.9, the test-retest reliability of the difference score should be restricted to a range of 0 to 0.5. Thus, the test-retest reliability of 0.94 reported by Wöstmann et al. for their flanker effect is extremely unlikely and may not easily replicate.

4.4.4. *Cross-task correlations using RT difference scores*

Researchers frequently want to establish the convergent validity between difference-score measures derived from distinct tasks that have presumably "isolated" a process of interest such as interference control. For example, do Simon effects and flanker effects measure the same domain-free ability to assert interference control? In IDRT terms we want to estimate the correlation between the task-specific processing time for the Simon incongruent condition ($\boldsymbol{\Delta}_{SimonI}$) and that for the flanker incongruent condition ($\boldsymbol{\Delta}_{FlankerI}$) by using the difference-scores (viz., the Simon and flanker effects) to achieve greater isolation of the processing devoted to conflict resolution. If both tasks recruit the same mechanism, then the underlying correlations ($\rho_{\Delta_{SimonI}\ \Delta_{FlankerI}}$) should be high.

How accurately does Corr[D_{Simon}, $D_{Flanker}$], the observed correlation between the Simon and Flanker effects, estimate/isolate $\rho_{\Delta_{SimonI}\ \Delta_{Flanker}I}$? The analyses provided by Miller and Ulrich using IDRT yield a complicated answer to that question. A first and general point is that the possible values of the target correlation ($\rho_{\Delta_{SimonI}\ \Delta_{FlankerI}}$) are tightly constrained by the correlations between other pairs of task-specific processing times (e.g., $\rho_{\Delta_{SimonC}\ \Delta_{SimonI}}, \rho_{\Delta_{FlankerC}\ \Delta_{FlankerI}}, \rho_{\Delta_{FlankerC}\ \Delta_{SimonC}}$). For example, if the underlying correlation between the two congruent conditions is zero (e.g.,$\rho_{\Delta_{FlankerC}\ \Delta_{SimonC}} = 0$) and if both difference scores involve opposing task-specific processes, then the target correlation must lie within the range of –0.1–0.3 as other values are impossible. Alternatively, when the congruent and incongruent conditions are more neutral than opposing a full range of observed correlations are possible. In our study the correlation between the Simon effect and flanker effect was significant, but quite small (r = 0.246, p = 0.029). Within the context and assumptions of the IDRT, the modest observed correlation of 0.246 may be almost as good as it gets if congruent and incongruent trials involve opposing task-specific processes. Or, that same correlation, may be viewed as quite far from the ideal if the central processing on congruent and incongruent trials is neutral rather than opposing.

In general, Miller and Ulrich conclude that the observable correlations involving RT means and/or difference scores depend on many factors influencing performance within the task. These include characteristics such as the difficulty and time needed to perform perceptual, central, and motor processing. The observable correlations also depend on both the variability of general and task-specific processing times and on the correlation between them. One can only agree with Miller and Ulrich that these findings motivate the need for extreme caution in interpreting observed correlations *"because there are cases in which correlations can be expected to be far higher or far lower than the correlations of the internal parameters that they might be intuitively assumed to measure"* p. 839. We add to this the observation that the implied presence of day effects makes it even more difficult to interpret the correlations between mean RT and/or RT difference scores. To this point day effects have been characterized as fluctuations in the participant's arousal, alertness, or motivation that are likely to change more from one day to another than from one trial to another. However, the internal parameters associated with task processing may also change as a result of learning and this too can adversely affect the reliability of a measure. This problem is explored in the next section.

5. Limitations associated with practice effects and speed-accuracy trade-offs

5.1. *Practice effects and reliability*

Classical test theory assumes that internal and external conditions are stable across time. But performance clearly improves with practice in choice RT tasks. A standard protocol employed in the present study is to filter out the early and dramatic practice effects by including a "practice" block before the experimental conditions. A small dose of practice, say 40 trials, helps considerably but performance on the flanker task continues to improve even after 100 sessions and a total of 20,000 trials (K. Paap et al., 2014). On the other hand, extended practice may eliminate the individual differences in true scores that a study seeks to detect. In this section our measures of inhibitory control are used to illustrate that adverse effects of practice can occur both within a session and between.

Table 1 provides the means and test-retest reliabilities of these measures across an average delay of one week. Two (antisaccade

[8] We thank Jeff Miller for guidance in the analysis presented above and for pointing out the importance of day effects.

Table 2
Block 1 to Block 2 reliability within each session.

Task Measure	T/C	Block 1		Block 2		r	SBP
		M1	SD 1	M2	SD 2		
Antisaccade							
RT Session 1	30	630	298	559	178	0.839**	0.912
RT Session 2	30	525	176	518	188	0.967**	0.983
RT Cost Session 1	30/30	79	161	8	101	0.142	0.249
RT Cost Session 2	30/30	15	88	8	94	0.863**	0.876
Flanker							
Inc. Session 1	32	574	59	574	60	0.858**	0.948
Inc. Session 2	32	538	50	540	52	0.932**	0.976
Con. Session 1	32	492	58	491	57	0.848**	0.943
Con. Session 2	32	470	49	460	49	0.900**	0.964
Global RT Session 1	64	533	57	533	56	0.870**	0.953
Global RT Session 2	64	504	48	500	49	0.937**	0.978
Inc – Con Session 1	32/32	82	25	83	31	0.595**	0.815
Inc – Con Session 2	32/32	69	23	81	23	0.543**	0.781
Simon							
Inc. Session 1	20	487	51	447	41	0.798**	0.888
Inc. Session 2	20	469	49	461	47	0.876**	0.934
Con. Session 1	20	454	46	440	42	0.770**	0.870
Con. Session 2	20	440	42	429	41	0.741**	0.851
Global RT Session 1	40	471	46	462	43	0.947**	0.973
Global RT Session 2	40	454	44	445	42	0.880**	0.936
Inc – Con Session 1	20/20	33	31	30	24	0.504**	0.670
Inc – Con Session 2	20/20	29	25	32	27	0.273*	0.546

Note: T/C = trials per condition, M 1 = mean block 1, M 2 = mean Block 2, SBP = Spearman-Brown prophecy, Inc = incongruent trials, Con = congruent trials.

and flanker) of the three tasks used to evaluate inhibitory control show that practice significantly reduces the interference score.

For antisaccade cost the Congruency × Session interaction was significant, F(1,74) = 7.66, p = 0.007, $\eta^2 = 0.094$. Practice reduced the antisaccade cost by 32 ms. For the flanker task the Congruency × Session interaction was significant, F(1,74) = 9.50, p = 0.003, $\eta^2 = 0.114$ with the flanker effect 8 ms smaller in Session 2. In contrast, the results for the Simon task show that a day's practice sped both trial types by the same amount. Despite a highly significant main effect of Session, F(1,74) = 24.50, p = 0.001, $\eta^2 = 0.249$ the Congruency × Session interaction was not significant, F(1,74) = 0.0.15, p = 0.700, $\eta = 0.002$. The magnitude of the Simon effect was reduced by only 1 ms in Session 2. The large differential practice effects across the neutral baseline and antisaccade conditions likely contribute to the modest test-reliably of the antisaccade cost measure.

Table 2 shows the mean and block-to-block correlations within each of the two sessions. As discussed earlier correlations between subsets of trials within a single session will be higher if fluctuations in attention, arousal, and arousal are more stable within a session than across days. However, differential practice effects across the two types of trials in a difference score can erode within session reliability as well. "Split-half" measures of reliability are usually based on a random partition of the "items", but the correlations shown in Table 2 are between the first two blocks of each session so that practice effects can be observed.

Inspection of the block to block correlations (column r) and Spearman-Brown prophecy values (column SBP) in Table 2 confirm that the reliabilities of the difference-score measures are usually higher within a single session than between sessions. One anomalous pattern is that the antisaccade cost measure has a block-to-block correlation of only 0.14 in Session 1, but it does blossom into a much higher correlation of 0.78 in Session 2. Despite the 15 trials of practice it is clear from inspection of the block means and standard deviations that there are marked improvements in performance from the first to second block on the first day. In contrast, inspection of performance on the antisaccade trials in Session 2 shows very little difference between the two blocks indicating that the effects of practice have stabilized. In turn, this enables the high block-to-block reliability in Session 2 for the antisaccade cost measure.

The remaining anomaly in the block-to-block correlations is the relatively small, but statistically significant correlation of 0.27 for the Simon effect (Inc – Con) in Session 2. This has nothing to do with practice effects within the session as the Simon effect differed by only 3 ms across the two blocks.

5.2. Effects of speed-accuracy trade-offs

As influentially described by Pachella (1974) observed RTs can be contaminated by speed-accuracy trade-offs. One approach is to use formal models that combine RT and accuracy into a single measure of processing efficiency (Ratcliff, 1978). A more modest strategy is to simply check for evidence of a trade-off. As reported overall accuracy for our flanker and Simon tasks was very high and that for the antisaccade task and color-shape switching task comfortably above 90% correct. It is therefore not surprising that the correlations between RT and PC are small and non-significant for the flanker and Simon tasks. The correlation between latency and accuracy on the antisaccade trials was r = −0.37 and for the switch trials r = −0.35. Thus, for our more difficult tasks there is no trade-off as individuals who are faster also tend to be more accurate.

Although there is no evidence for a speed-accuracy trade-off in our data, it is useful to consider the consequences when there is evidence of a trade-off does occur. The consequences for test-retest reliability may be benign if the strategy adopted by individuals remains stable when the instructions, error feedback, and rewards remain the same across the sessions. However, the consequences for evaluating the cognitive ability of a single individual are not benign as a person adopting a conservative strategy on an RT test is disadvantaged. One practical solution is to form a composite measure, for example, the efficiency score (ES) proposed by Townsend and Ashby (1983) that divides the mean RT by the proportion correct. There has been no formal analysis of efficiency scores comparable to Miller and Ulrich's RT analysis, but it appears that transforming RTs to ESs sometimes leads to a modest bump in test-retest reliability. For example, the test-retest reliability of the Simon RT effect 0.43 increases to a Simon ES effect of 0.50.

6. Summary and conclusions

With respect to our empirical contribution the test-retest correlations for several measures of EF derived from the Simon, flanker, antisaccade, and color-shape switching task were obtained in college students after an average delay of one week. The reliability of the RT measures based on a single mean were quite good, but somewhat lower for the incongruent trials in the Simon task. The test-retest correlations for the RT difference scores were statistically significant but ranged from only 0.43 for the Simon effect to 0.75 for mixing cost in the color-shape task. With the repeated caveat that specific instantiations of each task can be critical, our Simon task with 40 trials per condition (not an unusually low number) yields measures with unattractive levels of reliability. The test-retest reliability of our flanker task was somewhat higher, but far less than the surprisingly high reliability reported by Wöstmann et al. As discussed it is not easy to interpret or model test-retest reliability when there are strong day effects, as is apparently the case for both our flanker task and Wöstmann et al.'s flanker task.

Based on our discussion of the Miller and Ulrich IDRT model, the task impurity problem is perhaps worse than commonly feared for measures based on a single mean RT. Furthermore, taking a difference score does not unambiguously isolate a process of interest such as inhibitory control. Difference scores are preferable, but only if the specific instantiation of the task has demonstrated high levels of test-retest reliability. However, a high level of test-retest reliability for RT differences is the exception, not the rule, especially when the underlying correlation between congruent and incongruent conditions is high.

Assessments of convergent validity through cross-task correlations are constrained by the reliability of the measure derived from each task. The low to moderate levels of test-retest reliability obtained for most RT difference scores are contributing to the low levels of convergent validity between measures that were hypothesized to rely on the same central process (e.g., inhibitory control). But reliability should not shoulder all of the blame because Miller and Ulrich have shown that the underlying targeted processing can be moderately correlated across tasks and still yield low observed correlations. On the other hand, they also showed that moderate cross-task correlations can be the product of correlations in underlying "nuisance" parameters. From a psychometric perspective researchers looking at individual and group differences in EF may not be measuring what they hope to measure and far more work is needed to refine our models and tools and eventually lift the current fog when it comes to interpreting cross-task correlations.

Acknowledgements

We are deeply indebted to Jeff Miller for his patient tutoring on classical test theory and the Miller and Ulrich IDRT model. We thank the following past and present members of the LACE (Language, Attention, & Cognitive Engineering) laboratory for their individual contributions to this project: Edgar Alcaine, Kirstin Anderson, Morgan Bockelman, Jack Darrow, Frank Du, Richard Farrell, Lynne Freeman, Zack Greenberg, Hunter (Johnson) Myuz, Yunyun Liu, and Ariz Sanchez.

References

Bakker, M., van Dijk, A., Wicherts, J.M., 2012. The rules of the game called psychological science. Perspect. Psychol. Sci. 7, 534–554.

Burgess, P.W., 1997. Theory and methodology in executive function research. In: Rabbitt, P. (Ed.), Methodology of Frontal and Executive Function. Psychology Press, Hove, pp. 81–111.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. 14, 365–376.

Callejas, A., Lupianez, J., Funes, M.J., Tudela, P., 2005. Modulations among the alerting orienting: and executive control networks. Exp. Brain Res. 167, 27–37.

Costa, A., Hernández, M., Sebastián-Gallés, N., 2008. Bilingualism aids conflict resolution: evidence from the ANT task. Cognition 106, 59–86.

Delis, D.C., Kaplan, E., Kramer, J.H., 2001. Delis-Kaplan Executive Function System (D-KEFS). The Psychological Corporation, San Antonio, TX.

Egner, T., 2008. Multiple conflict-driven control mechanisms in the human brain. Trends Cogn. Sci. 12, 374–380.

Eriksen, B.A., Eriksen, C.W., 1974. Effect of noise letters upon the identification of a target letter in a nonsearch task. Percept. Psychophys. 16, 143–149.

Fan, J., McCandliss, B.D., Sommer, T., Raz, A., Posner, M.I., 2002. Testing the efficiency and independence of attentional networks. J. Cogn. Neurosci. 14, 340–347.

Gualtieri, C.T., Johnson, L.G., 2006. Reliability and validity of a computerized neurocognitive test battery: CNS Vital Signs. Arch. Clin. Neuropsychol. 21, 623–643.

Hilchey, M.D., Klein, R.M., 2011. Are there bilingual advantages on nonlinguistic interference tasks? Implications for plasticity of executive control processes. Psychon. Bull. Rev. 18, 625–658.

Hilchey, M.D., Saint-Aubin, J., Klein, R.M., 2015. Does bilingual exercise enhance cognitive fitness in non-linguistic executive processing tasks. In: Schwieter, J.W. (Ed.), Cambridge Handbook of Bilingual Processing. Cambridge University Press.

Ishigami, Y., Klein, R.M., 2010. Repeated measurement of the components of attention using two versions of the attention network test (ANT): stability, insolubility, robustness, and reliability. J. Neurosci. Methods 190, 117–128.

Ishigami, Y., Klein, R.M., 2011. Repeated measurement of the components of attention of older adults using the two versions of the Attention Network Test: stability, isolability, robustness, and reliability. Front. Aging Neurosci. 3, 1–17 (Article 17).

Paap, K., Wagner, S., Johnson, H., Bockelman, M., Cushing, D., Sawi, O., 2014 May. 20,000 flanker task trials: Are the effects stable reliable, robust, and stable? Poster presented at the Association for Psychological Science meeting. San Francisco, CA.

Paap, K.R., Johnson, H., Sawi, O., 2014. Are bilingual advantages dependent upon specific tasks or specific bilingual experiences? J. Cogn. Psychol. 26 (6), 615–639.

Kane, M.J., Bleckley, M.K., Conway, A.R.A., Engle, R.W., 2001. A controlled-attention view of working-memory capacity. J. Exp. Psychol. Gen. 130 (2), 169–183.

Klein, C., Fischer, B., 2005. Instrumental and test-retest reliability of saccadic measures. Biol. Psychol. 68, 201–213.

Kornblum, S., 1994. The way irrelevant dimensions are processed depends on what they overlap with: the case of Stroop- and Simon-like stimuli. Psychol. Res. 56, 130–135.

Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Reading: Addison-Wesley.

MacLeod, J.W., Lawrence, M.A., McConnell, M.M., Eskes, G.A., Klein, R.M., 2010. Appraising the ANT: Psychometric and theoretical considerations of the attention network test. Neuropsychology 24 (5), 637–651.

Miller, J., Ulrich, R., 2013. Mental chronometry and individual differences: modeling reliabilities and correlations of reaction time means and effect sizes. Psychon. Bull. Rev. 20, 819–858.

Miyake, A., Friedman, N.P., 2012. The nature and organization of individual differences in executive functions: four general conclusions. Curr. Direct. Psychol. 21 (1), 8–14.

Monsell, S., Mizon, G.A., 2006. Can the task-cuing paradigm measure an endogenous task-set reconfiguration process? J. Exp. Psychol. 32, 493–516.

Paap, K.R., Greenberg, Z.I., 2013. There is no coherent evidence for a bilingual advantage in executive processing. Cognit. Psychol. 66, 232–258.

Paap, K.R., Sawi, O., 2014. Bilingual advantages in executive functioning: problems in convergent validity, divergent validity, and the identification of the theoretical constructs. Front. Psychol. 5, 1–15 (962).

Paap, K.R., Johnson, H.A., Sawi, O., 2015. Bilingual advantages in executive functioning either do not exist or are restricted to very specific and undetermined circumstances. Cortex 69, 265–278.

Paap, K.R., Myuz, H.A., Anders, R.T., Bockelman, M.F., Mikulinsky, R., Sawi, O., 2016. No compelling evidence for a bilingual advantage in switching or that frequent language switching reduces switch cost. J. Cogn. Psychol.

Paap, K.R., 2014. The role of componential analysis, categorical hypothesizing, replicability and confirmation bias in testing for bilingual advantages in executive functioning. J. Cogn. Psychol. 26 (3), 242–255.

Pachella, 1974. The interpretation of reaction time in information processing research. In: Kantowitz, B. (Ed.), Human Information Processing: Tutorials in Performance and Cognition. Halstead Press, New York, pp. 41–82.

Prior, A., MacWhinney, B., 2010. A bilingual advantage in task switching. Bilingualism 13, 253–262.

Ratcliff, R., 1978. A theory of memory retrieval. Psychol. Rev. 85, 59–108.

Rubin, O., Meiran, N., 2005. On the origins of the task mixing cost in the cued switching paradigm. J. Exp. Psychol. 31, 1477–1491.

Salthouse, T.A., 2010. Is flanker-based inhibition related to age? Identifying specific influence of individual differences on neurocognitive variables. Brain Cogn. 73, 51061.

Stoffels, E.J., van der Molen, M.W., 1988. Immediate arousal and location effects of auditory noise on visual choice reaction time. Percept. Psychophys. 44 (1), 7–14.

Townsend, J.T., Ashby, F.G., 1983. The Stochastic Modeling of Elementary Psychological Processes. Cambridge University Press, Cambridge.

Unsworth, N., Spillers, G.J., 2010. Working memory capacity: attention control, secondary memory, or both? A direct test of the dual-component model. J. Memory Lang. 62, 392–406.

Unsworth, N., McMillan, B.D., Brewer, G.A., Spillers, G.J., 2012. Everyday attention failures: an individual differences investigation. J. Exp. Psychol. 38, 1765–1772.

Unsworth, N., Fukuda, K., Awh, E., Vogel, E.K., 2014. Working memory and fluid intelligence: capacity, attention control, and secondary memory retrieval. Cognit. Psychol. 71, 1–26.

Wöstmann, N.M., Aichert, D.S., Costa, A., Rubia, K., Möller, H.J., Ettinger, U., 2013. Reliability and plasticity of response inhibition and interference control. Brain Cogn. 81, 82–94.

Update

Journal of Neuroscience Methods

Volume 307, Issue , 1 September 2018, Page 260

DOI: https://doi.org/10.1016/j.jneumeth.2018.07.018

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Corrigendum

Corrigendum to “The role of test-retest reliability in measuring individual and group differences” [Journal of Neuroscience Methods 274 (2016) 81–93]

Kenneth Paap[a,*], Oliver Sawi[b]

San Francisco State University
University of Connecticut

On page 87 we stated that “our Day 1 flanker effect (M = 83), D = 26) was greater than that of Wöstman et al.’s (M = 32, SD = 32).” he correct mean for Wöstmann et al.’s flanker effect was M = 58.5 ms not 32 ms as we reported. The authors would like to apologize for any inconvenience caused.

DOI of original article: https://doi.org/10.1016/j.jneumeth.2016.10.002

* Corresponding author.
E-mail address: kenp@sfsu.edu (K. Paap).

ttps://doi.org/10.1016/j.jneumeth.2018.07.018