Publisher: Taylor & Francis & The Experimental Psychology Society

Journal: *The Quarterly Journal of Experimental Psychology*

DOI: 10.1080/17470218.2016.1239750

Inhibition in Task Switching:

The Reliability of the N–2 Repetition Cost

Agnieszka W. Kowalczyk and James A. Grange

School of Psychology, Keele University, UK

Word Count: 7,153 (Main body, not including abstract or reference list)

Author Note

Please address correspondence to Agnieszka W. Kowalczyk, School of Psychology, Dorothy Hodgkin Building, Keele University, Keele, ST5 5BG. Email: a.w.kowalczyk@keele.ac.uk. All raw data and analysis code are available to download at http://bit.ly/29fhj9u

Abstract

The n–2 repetition cost seen in task switching is the effect of slower response times performing a recently completed task (e.g. an ABA sequence) compared to performing a task that was not recently completed (e.g. a CBA sequence). This cost is thought to reflect cognitive inhibition of task representations and has been well replicated (Koch, Gade, Schuch, & Philipp, 2010). As such, the n–2 repetition cost has started to be used as a measure of individual differences in inhibitory control (e.g. Whitmer & Banich, 2007); however, the reliability of this measure has not been investigated in a systematic manner. The current study addressed this important issue. Seventy-two participants performed three task switching

paradigms; participants were also assessed on rumination traits and processing speed—measures of individual differences potentially modulating the n–2 repetition cost. We found significant n–2 repetition costs for each paradigm. However, split-half reliability tests revealed that this cost was not reliable at the individual-difference level. Neither rumination tendencies nor processing speed predicted this cost. We conclude that the n–2 repetition cost is not reliable as a measure of individual differences in inhibitory control.

Keywords: task switching, n–2 repetition cost, backward inhibition, reliability

Inhibition in Task Switching: The Reliability of the N–2 Repetition Cost

Our ability to perform efficient task switching can be taken for granted, given the ease with which we are able to do it; for example, when writing a manuscript, one can alternate between writing and reading previously prepared notes, answering the phone, and surfing the internet. However, what may seem an effortless behaviour arises because of mental processes working together, ultimately resulting in humans being able to behave in a goal-directed and context-appropriate manner. For goal-oriented behaviour to be effective, one has to be able to attend to relevant information and ignore irrelevant information (e.g. ignoring social media notifications when working on a manuscript). The mental processes coordinating the ability to maintain context-relevant behaviour and change it when necessary are not fully understood, but it seems they are part of a dynamic system (Goschke, 2000), a system by which behaviour is adapted in a moment-to-moment manner, by activating relevant- and inhibiting irrelevant-dimensions of a given task (Mayr & Keele, 2000).

In the laboratory, researchers use the so-called task switching paradigm to assess the efficiency with which switching is performed; data from these paradigms (e.g. reaction times, RT; accuracy) are used to make inferences about candidate mental processes associated with switching (see Grange & Houghton, 2014; Kiesel et al., 2010; and Vandierendonck,

Liefooghe, & Verbruggen, 2010, for recent reviews on task switching). Task switching paradigms typically require participants to make rapid responses to stimuli presented sequentially. For example, participants might be presented with numerical stimuli and be asked to perform tasks such as judging whether the number is odd/even (a parity judgment), lower/higher than 5 (a magnitude judgment), or whether it is in red/blue font (a colour judgment).

One cognitive process thought to aid task switching performance is the inhibition of recently performed task-sets (i.e., the mental representation required to perform a task; Mayr & Keele, 2000). Evidence for inhibition in task switching comes from the backward inhibition paradigm, where participants are required to switch between three tasks. Experiments using this paradigm show that people are slower and less accurate performing a task that was performed recently (i.e., an ABA sequence) compared to performing a task that was not performed recently (i.e., a CBA sequence; where A, B, and C are arbitrary labels for tasks). This effect—known as the *n–2 repetition cost*—is interpreted as evidence for inhibitory control (Gade, Schuch, Druey, & Koch, 2014; Koch et al., 2010; Mayr & Keele 2000): It is thought to reflect the persisting inhibition applied to task A when it was switched away from in preference for task B in an ABA sequence. The n–2 repetition cost has been replicated in a number of different studies and—to date—seems resistant to non–inhibitory accounts (Koch et al., 2010; Mayr, 2007).

Much is known about the characteristics of the n–2 repetition cost and how it relates to cognitive inhibition (see Gade et al., 2014, for a recent review). As such, the paradigm has become of interest to researchers wishing to explore cognitive inhibition more widely. For example, researchers have used the n–2 repetition cost to assess inhibitory control from a variety of approaches, including the effect of brain lesions (Mayr, Diedrichsen, Ivry, & Keele, 2006), neuroimaging (Dreher, Kohn, & Berman, 2001; Whitmer & Banich, 2012),

healthy ageing (Lawo, Philipp, Schuch, & Koch, 2012; Mayr, 2001), Parkinson's disease (Fales et al., 2006); Williams' syndrome (Foti et al., 2015), obsessive-compulsive disorder (Moritz, Hübner, & Kluwe, 2004), major-depressive disorder (Whitmer & Banich, 2012), pathological gambling (Yiu-kwan, 2008), bilingualism (Philipp & Koch, 2009; Prior, 2012), and mindfulness (Greenberg, Reiner, & Meiran, 2013).

The n–2 repetition cost has also been utilised to assess individual differences in inhibitory control by relating it to other variables. For example, Whitmer and Banich (2007) found a negative correlation between the n–2 repetition cost and the tendency to engage in depressive rumination. In unpublished work, Grange (2010) found no correlation between n–2 repetition costs and working memory capacity, as measured by the automated operation-span task (Unsworth, Heitz, & Engle, 2005). In another study—although not their primary focus—Grange and Juvina (2015) presented data from individual subjects on the effect of practice on the n–2 repetition cost. These data showed considerable within- and between-subject variation in the magnitude of the n–2 repetition cost and its reduction with practice.

The Current Study

As the n–2 repetition cost garners wider attention as a tool to study cognitive inhibition, it becomes important to assess its psychometric properties (Drost, 2011; Onwuegbuzie & Daniel, 2002). Although well replicated, no one has yet provided a systematic assessment of the internal reliability of the n–2 repetition cost; we sought to provide some information regarding this in the current study. Some measures commonly used to assess individual differences in cognitive/inhibitory control have been shown to be reliable (stop-signal task: Congdon et al., 2012; go/no-go task: Leue, Klein, Lange, & Beauducel, 2013; Stroop test: Strauss, Allen, Jorgensen, & Cramer, 2005). However, other measures appear to have low reliability (e.g., the negative priming effect; Bestgen & Dupont, 2000).

To our knowledge, there is only one study that has partially addressed the reliability of the n–2 repetition cost; Pettigrew and Martin (2015)—amongst other results—reported the reliability of the n–2 repetition cost as low (Spearman-Brown corrected correlation coefficient .44–.51). For a given measure to be considered reliable, its Spearman-Brown correlation coefficient must be at least .7 (Cronbach, 1951). The finding that the n–2 repetition cost has a low reliability is an indication that this effect should be interpreted with caution when used as a measure of individual differences in inhibitory control (e.g., Whitmer & Banich, 2007).

In the current study, we wished to examine in a systematic manner the reliability of the n–2 repetition cost. Participants performed three versions of the task switching paradigm, similar to paradigms used in published research on the n–2 repetition cost: The "Target Detection" paradigm (similar to Houghton, Pritchard, & Grange, 2009, Experiment 3) the "Visual Judgment" paradigm (similar to Gade & Koch, 2008); and the "Numeric Judgment" paradigm (similar to Schuch & Koch, 2003). These paradigms differed according to task cues, stimuli, and response requirements, thus allowing some generalisations of the findings. Exposing participants to three paradigms allowed us to explore the internal reliability of each paradigm, but also to explore for the first time the correlation of n–2 repetition cost between paradigms. We also measured depressive rumination (shown to modulate the n–2 repetition cost; Whitmer & Banich, 2007) and general processing speed as potential controls during the reliability analysis. To anticipate the results, we find that the n–2 repetition cost across all three paradigms has very low internal reliability.

Methods

Participants

The Ethical Research Panel at Keele University approved the study. Participants were first year Psychology students from Keele University, and participated in exchange for partial course credit. Participants were required to be at least 18 years old, understand spoken and written English, and to have normal or corrected-to-normal vision.

Our sample size was determined using the R package "pwr" (Champely, 2009), using the expected effect size of the reliability measure. As explained later, we used a form of split-half reliability to assess reliability. The criterion for reliability relates to a Pearson product-moment correlation coefficient of $r \approx .54$. Using this expected effect size, and a desired power of 95%, the required sample size was 38. However, to be conservative, we reduced our expected correlation coefficient to $r = 0.4$, which for power of 95% requires 75 participants. We used this as our intended sample size.

Ninety-four participants were recruited. Twenty-two participants were removed: fourteen due to accuracy being lower than an a priori defined criterion of 90% in at least one of the task switching paradigms; seven due to incomplete data (attending only one session out of two); and one due to unusually large n–2 repetition *benefit* (> 700 milliseconds, ms). The final sample consisted of 72 participants (60 females; mean age = 18.76, SD = 1.07). Note this is three below our intended sample size, but still maintains more than 94% power.

General Procedure

Participants attended two sessions each lasting 45 minutes (1–8 days apart, $M = 3.10$, $SD = 2.30$) during which they performed three task switching paradigms, a processing speed task, and filled in a rumination tendencies-questionnaire. The order of all of the components was counterbalanced across participants. Each session started either with an administration of the questionnaire or the processing speed task (which alternated across participants); the three task switching paradigms were counterbalanced with a Latin Square Design (3x3). The task

switching paradigms were presented on a standard PC with a 17in. monitor via E-Prime v. 2.0 software (Psychology Software Tools, Pittsburgh, PA). Responses were made on a 1-ms precise USB keyboard.

Before performing each task switching paradigm, participants had a practice consisting of 16 trials, which was repeated once if four or more errors were made. No further practice was allowed. During the practice, if an incorrect response was made, the word 'Error' (font the Courier New, size 18) appeared on the screen for 1000ms.

The trial structure for each task switching paradigm was the same. First, a cue was presented for 500ms, followed by the stimulus that stayed on the screen for the duration it took participants to make a response. Participants learned the cue–task pairings before the practice phase. The cue remained on the screen during stimulus presentation. Once a response was recorded, the cue for the next trial appeared 100ms later (response–cue interval). Participants were asked to respond as quickly and as accurately as possible. Participants performed four blocks of 120 trials each. However, due to a coding error, the Target Detection paradigm only presented 360 trials[1]. Trials were presented in a random order with the constraint that immediate task repetitions were not allowed; this was because immediate repetitions decrease the magnitude of the n–2 repetition cost (Philipp & Koch, 2006). Trials were classified as ABA or CBA sequences by comparing the current task requirement to that at n–2. An overview of the trial structure for each paradigm is shown in Figure 1.

Insert Figure 1 Here

Task Switching Paradigms

[1]Additional analysis limited to the first 360 trials per paradigm was also conducted, to examine to what extent unequal number of trials in task switching paradigms affected the results. Differences from additional analyses are reported in footnotes.

The Target Detection paradigm. This paradigm was based on the procedure of Houghton, Pritchard, and Grange (2009), Experiment 3, and required participants to make spatial localisation judgments of stimuli (Mayr & Keele, 2000).

The cues were shapes (triangle, square, octagon; height and width 4 cm) and targets were oval shapes of different characteristics (see Figure 1). All ovals had a height of 6 cm; three ovals had widths of 2.3 cm, and one oval had a width of 3.5 cm. Cues and stimuli were presented in grayscale shading on a white background. Participants were required to respond to the location of the target associated with the presented cue. For all participants, the square cue was associated with the "shaded" target, the triangle cue was associated with the "bordered" target, and the octagon cue was associated with the "angled" target. The cue appeared in the centre of the screen, followed by four oval shapes (three possible targets and one distractor), with one oval centred within each quadrant of the screen. The position of the targets on the stimulus display was randomised.

Participants responded to the location of the correct target by pressing one of four possible response keys, depending on the location of the correct target. Each key corresponded to one corner of the screen: upper left-D, lower left-C, lower right-N, upper right-J. Participants used their index and middle fingers of both hands to respond with; middle fingers on D/J, and index fingers on C/N.

The Visual Judgments paradigm. This paradigm is based on the procedure of Gade and Koch (2008). Participants made judgments about the visual characteristic of a single multivalent stimulus. The stimulus on each trial was either the letter "A" or the number "4", in either red or blue font; it could also be large or small. Participants were required to judge whether the stimulus was a digit or a number (a form task), small or large (0.5cm vs. 1cm; a size task), or whether it was red or blue (a colour task).

The stimulus appeared in the centre of a white rectangle (4 cm high x 3.5 cm wide) and a task was cued by four cues of one type presented around the rectangle; each cue centred to each side of the rectangle. The cue was the $ sign (1 cm high) for the judgment of 'form', arrows pointing up and down (1 cm high) for the judgment of 'size', and yellow squares (1 x 1 cm) for the judgement of the colour task. Participants responded with their index fingers by pressing one of two keys: "Z" for a "blue", "small", or "letter" response, and "M" for a "red", "large", or "digit" response.

The Numeric Judgments paradigm. This paradigm used the type of stimuli previously used by Schuch and Koch (2003), with central/peripheral judgment replaced by the word/digit judgment. The stimulus presented on each trial was either a digit (1, 2 ,3 ,4 ,6 ,7 ,8 ,9) or a number word (one, two, three, four, six, seven, eight, nine). Participants were required to judge whether the stimulus was odd or even (a parity judgment), whether it was lower or higher than five (a magnitude judgment), or judge whether the stimulus was in digit or word form (a form judgment). Task cues were the words "parity", "magnitude", and "form". Both, the cue and the stimulus were presented on the screen in black, Courier New, size 24 font on a grey background. The cue was presented above a central fixation cross, and the stimulus was presented below fixation (see Figure 1). Participants responded with their index fingers by pressing one of two keys: "Z" for an "odd", "lower than five", or "word" response, and "M" for an "even", "higher than five", or "digit" response.

Materials

The processing speed task. The processing speed task was used to assess the rate at which participants processed information; the test used was an adaptation of the Digit Symbol Substitution Task (e.g. van der Elst, van Boxtel, van Breukelen, & Jolles, 2006). This factor has not been directly linked to the n–2 repetition cost, but processing speed is

known to predict overall cognitive abilities (Brown, Brockmole, Gow, & Deary, 2012; Stawski, Sliwinski, & Hofer, 2013), so we wished to potentially control for it.

Participants were presented with nine digit-letter pairs on a sheet of paper; beneath the number-letter pairings was a list of digits; participants were required to write the letter that corresponded to each digit. Participants were given 120 seconds to match as many pairs of letters with numbers as possible. Participants practiced this task with seven pairs; after the practice they matched the rest of the pairs in a sequential manner, without skipping any numbers; they could correct a mistake but could not erase their answers (responses were written with a pencil). The total score is the number of correct matches (maximum 133), with higher scores reflecting more efficient processing speed.

The Ruminative Response Scale (RRS). Ruminative tendencies were assessed with the Ruminative Response Scale (RRS) – short version, consisting of ten questions, including the Brooding and Reflection parts (Treynor, Gonzalez, & Nolen-Hoeksema, 2003). This version has been shown to be reliable (Nolen-Hoeksema & Morrow, 1991; Thanoi & Klainin-Yobas, 2015; Whitmer & Banich, 2007).

Participants answered the following question "…*how often you do things described in each statement*" in relation to ten different statements (e.g. *Think 'What I am doing to deserve this'*?). Participants responded as to what extent they felt the statements applied to them, using a 1–4 scale, where 1 corresponded to "almost never" and 4 corresponded to "almost always". The RRS scores were obtained by summing the answers participants circled; the maximum score was 40, and the minimum score was 10. Higher scores reflect stronger rumination tendencies.

Design

The current study employed a within-subjects design. The dependent variables were RTs and accuracy, and the independent variables were task *sequence* (ABA vs. CBA) and *paradigm* (Target Detection vs. Visual Judgment vs. Numeric Judgment). To perform regression between n–2 repetition cost, depressive rumination, and processing speed, predictors were the RRS questionnaire and processing speed test scores and the dependent variable was n–2 repetition cost. Details of the reliability procedure are outlined in the Results.

Results

Data Trimming

For the task switching data, we removed the first two trials of each block. For response time analysis, we removed error trials and the two trials following an error; this trimming led to 6% trials being removed (4.6% from the Target Detection paradigm, 5.7% from the Visual Judgment paradigm, and 7.4% from the Numeric Judgment paradigm). Both of these removals were due to the inability to classify the current trial as an ABA or CBA sequence. After the error trimming, we further trimmed response times by removing RTs faster than 150ms, and RTs slower than 2.5 standard deviations above each participant's mean for each cell of the experimental design. In total, the accuracy and RT trimming led to 11.7 % of trials being removed (9.9 % for the Target Detection paradigm, 11.3% for the Visual Judgment paradigm, and 13.7% for the Numeric Judgment paradigm).

Standard N–2 Repetition Cost Analysis

Before conducting the reliability analysis, we performed standard n–2 repetition cost analysis to assess whether the n–2 repetition cost was present in each paradigm. Additionally, we took this opportunity to assess whether the magnitude of the n–2 repetition cost varies reliably across the different paradigms. Mean response times and accuracy for each level of the design are presented in Table 1.

Insert Table 1 here

Response time analysis. We submitted the response times to a 2x3 repeated measures analysis of variance (ANOVA). There was a main effect of *sequence*, $F(1, 71) = 108.98$, $p<.001$, $\eta_g^2 = .02$, as participants were slower performing ABA sequences ($M = 1212$, $SE = 102$) compared to CBA sequences ($M = 1120$, $SE = 98$). There was a main effect of *paradigm*, $F(2, 142) = 61.19$, $p<.001$, $\eta_g^2 = .16$, as participants were fastest performing the Target Detection paradigm ($M = 968$, $SE = 76$) followed by the Visual Judgement paradigm ($M = 1230$, $SE = 103$), and the Numeric Judgement paradigm ($M = 1276$, $SE = 113$). There was no interaction between *sequence* and *paradigm*, $F(2, 142) = 0.67$, $p=.51$, $\eta_g^2 < .001$, suggesting equivalent n–2 repetition costs across the three paradigms. The n–2 repetition cost was significant for the Target Detection paradigm, $t(71) = 8.50$, $p<.001$, 95%CI = [69, 111], for the Visual Judgement paradigm, $t(71) = 9.12$, $p<.001$, 95%CI = [82, 128], and for the Numeric Judgment paradigm, $t(71) = 5.31$, $p<.001$, 95%CI = [54, 119]. See Figure 2 for density functions of the distributions of n–2 repetition cost across each paradigm for response times.

Insert Figure 2 here

Accuracy analysis. We submitted accuracy to a 2x3 repeated measures ANOVA. There was a main effect of *sequence*, $F(1,71) = 10.22$, $p<.01$, $\eta_g^2 = .01$, as participants were more accurate in CBA sequences (M = 97.22, SE = 0.27) compared to ABA sequences (M = 96.80, SE = 0.25). There was also a main effect of *paradigm*, $F(2, 142) = 19.07$, $p<.001$, $\eta_g^2 = .07$, as participants were most accurate performing the Target Detection paradigm (M = 97.66, SE = 0.22) followed by the Visual Judgment paradigm (M = 97.10, SE = 0.24), and the Numeric Judgment paradigm (M = 96.27, SE = 0.30). There was also an interaction between *sequence* and the *paradigm*, $F(2,142) = 4.01$, $p<.05$, $\eta_g^2 = .008$. The n–2 repetition cost was significant for the Target Detection paradigm, $t(71) = -4.71$, $p<.001$, 95%CI = [-1.31, -0.53], but it was not significant for the Visual Judgement paradigm, $t(71) = -1.42$, $p=.16$, 95%CI = [-0.78, 0.13] or for the Numeric Judgment paradigm, $t(71) = -0.01$, $p=.99$, 95%CI = [-0.52, 0.52]. See Figure 3 for density functions of the distributions of n–2 repetition cost across each paradigm for accuracy.

*** Insert Figure 3 here***

Overall Performance Correlations

We first performed Pearson product-moment correlation analysis on overall (i.e., mean) RT and accuracy to assess whether overall participant performance was stable across all three paradigms. These overall measures were also used as input into some of the individual differences analysis reported below, so we report them here for completeness.

Reaction time. The mean RT for the Target Detection paradigm correlated significantly with the Numeric Judgment paradigm ($r = .64$, $p< .001$) and the Visual Judgment paradigm ($r = .68$, $p <.001$); the Numeric Judgment paradigm correlated with the

Visual Judgment paradigm (r = .75, p <.001). All of these correlations remain significant when controlling for multiple comparisons (see Table 2).

Accuracy. Overall accuracy on the Target Detection paradigm correlated significantly with the Numeric Judgment (r = .42, p<.001), and the Visual Judgment paradigm (r = .50, p < .001); the Numeric Judgment correlated significantly with the Visual Judgment paradigm (r = .63, p < .001). All of these correlations remain significant when controlling for multiple comparisons (see Table 3).

Insert Table 2 here

Insert Table 3 here

Individual Differences Analysis

In this section, we report those analyses involving processing speed and rumination measures. The mean score on the processing speed test was 91.41 (SD = 12.78, min. = 66, max. = 121)[2], and the mean score on the RRS questionnaire was 19.04 (SD = 5.00, min. = 11, max. = 37). These two measures did not correlate (Table 2).

The mean RTs on the three paradigms were significantly negatively correlated with the processing speed score (Numeric Judgment, r = -.49, p<.001; Target Detection, r = -0.63, p< .001; the Visual Judgment, r = -.49, p < .001). This presents a form of manipulation check of our measures, as it suggests our measure of processing speed was successful. The RRS score did not correlate with the mean RTs on the three task switching paradigms.

[2] One participant had missing data for the Processing Speed measure; to maintain power, we kept this participant and imputed their value using the mean score for the Processing Speed test. Removing this participant changes none of the conclusions.

Multiple regression analyses revealed that the response time n–2 repetition cost was not predicted by either the RRS score or the processing speed score in any of the three task switching paradigms (see Table 4)[3]. As n–2 repetition cost was not predicted by the individual differences, we did not control for these measures in the between-paradigm correlations or the reliability analysis reported below.

Insert Table 4 here

Between-Paradigm Correlations of the N–2 Repetition Cost

In the next phase of analysis, we wished to assess whether measures of the n–2 repetition cost correlated between different task switching paradigms. To achieve this, we performed Pearson product-moment correlations on the n–2 repetition costs across all three paradigms separately for response times and accuracy. See Table 5 for the response time correlations, and Table 6 for the accuracy correlations.

Response time. The Target Detection paradigm did not correlate with the Visual Judgment paradigm (r = .07, p = .57), but it correlated with the Numeric Judgment paradigm (r = .25, p = .036); the Visual Judgment paradigm correlated with the Numeric Judgment paradigm (r = .30, p=.01; see Table 5)[4]. Note, though, that these do not remain significant when using Bonferroni corrections for multiple correlations.

Insert Table 5 here

[3]Transforming RRS and the processing speed scores into z-scores and then inputting them into the regression analysis also yielded non-significant results.

[4] Equal trials correlation analysis: the n–2 repetition cost correlated only between the Visual and Numeric Judgments paradigms (r = .28, p = .01). The reported correlations remained unchanged when we controlled for individual differences in processing speed via partial correlations: Target Detection paradigm did not correlate with the Visual Judgement paradigm (r = .07, p = .56), but it did correlate with the Numeric Judgement paradigm (r = .25, p = .03); the Visual Judgement paradigm correlated with the Numeric Judgement paradigm (r = .30, p = .01). Note these latter correlations do not remain significant when using Bonferroni corrections for multiple correlations.

Accuracy. The Target Detection paradigm did not correlate with the Numeric Judgment (r = .06, p = .64), or the Visual Judgment paradigm (r = -.15, p = .22); also, the Numeric Judgment paradigm did not correlate with the Visual Judgment paradigm (r = .01, p =.93; see Table 6).

Insert Table 6 here

Reliability Analysis

We conducted split-half reliability analysis to assess the reliability of the n–2 repetition cost in response times and accuracy. Unlike the test–retest procedure, the split-half method controls for practice effects, which is important for examining the n–2 repetition cost because it has been shown that practice reduces the n–2 repetition cost (Grange & Juvina, 2015). It is also less time consuming, and it is therefore suitable for testing the reliability of cognitive tests (Drost, 2011).

One potential disadvantage of the split-half method is that the method of split is often arbitrary (for example, splitting trials into odd and even numbered trials, and assessing the reliability of the DV between each half). The resulting reliability statistic is a point-estimate (i.e., a single value), and therefore there is no way of being sure whether the point estimate is an accurate estimate of the measure's reliability, or whether it is specific to the splitting method used.

To overcome this potential disadvantage, we performed a form of bootstrapping by conducting many random splits of the data and calculating a reliability estimate for each split (see e.g. Congdon et al., 2012). Specifically, for each paradigm and each participant, post-trimming data were split randomly into two halves. Then, for each half, mean RTs for ABA and CBA sequences were calculated. This allowed us to calculate the n–2 repetition cost for

each half. Next, a Pearson product-moment correlation between the n–2 repetition cost from the two halves was conducted, and the point-estimate was stored. We repeated this procedure 500 times, allowing for a distribution of correlation coefficients.

Splitting the data in half reduces the total number of data points being used in the reliability analysis, which can reduce the reliability coefficient. Therefore, it is typical to use the Spearman-Brown correction, which is given by

$$r_c = \frac{Nr}{1 + (N-1)r} \qquad (1)$$

where r is the Pearson product-moment coefficient and N is the number of "tests" being combined. In our case, we are combing two halves, so $N = 2$. The reliability of a given measure is considered as strong if r_c is at least .7 (Cronbach, 1951; Picardi & Masick, 2013; Revelle & Condon, 2014). With N = 2, this pertains to an uncorrected $r \approx .5385$.

Note that in our bootstrapping method, some (uncorrected) r values were negative, which indicates total lack of reliability, and renders the Spearman-Brown correction uninterpretable. As such, we report the uncorrected r below for the bootstrapping, but refer to r_c when interpreting the full result.

Response times. Figure 4 shows the reliability tests for the n–2 repetition cost in response times; it depicts a violin plot of the distribution of correlation coefficients for the split-half reliability for each of the three paradigms. A violin plot is like a standard box-plot, with the addition of a rotated density function of the distribution of scores, allowing a better description of the shape of the distributions.

Insert Figure 4 here

As can be seen, the peaks of all of the distributions fall short of the criterion for reliability (as stated, equivalent to an uncorrected $r \approx .5385$); whilst the tails of the distributions for the Target Detection and the Numeric Judgement paradigm do cross the criterion, we do not consider this strong evidence for acceptable reliability. The median values of uncorrected correlation coefficients were: $r = .35$ for the Target Detection paradigm, $r = .23$ for the Visual Judgment paradigm, and $r = .43$ for the Numeric Judgment paradigm. These translate to corrected values of $r_c = .52$ for the Target Detection paradigm, $r_c = .37$ for the Visual Judgement paradigm, and $r_c = .60$ for the Numeric Judgement paradigm.

Accuracy. Figure 5 shows the reliability tests for the n–2 repetition cost in the accuracy data. For all three paradigms, the whole of the reliability distribution is beneath the criterion for reliability. The median values of uncorrected correlation coefficients were: $r = .07$ for the Target Detection paradigm, $r = .18$ for the Visual Judgment paradigm, and $r = .19$ for the Numeric Judgment paradigm. These translate to corrected values of $r_c = .13$ for the Target Detection paradigm, $r_c = .31$ for the Visual Judgement paradigm, and $r_c = .32$ for the Numeric Judgement paradigm.

Insert Figure 5 here

Exploratory analysis of the n–2 repetition cost: Practice effect. It could be argued that due to a relatively short practice period for each of our task switching paradigms, participant performance had not reached asymptote before the main experimental blocks commenced. If this is the case, the lack of reliability we have observed could be due to our

data not being reflective of optimum performance from participants[5]. In order to rule out this possibility, we conducted exploratory analysis where for each of the paradigms the first block of the data was removed before the split-half bootstrapping was conducted. To anticipate, the results were qualitatively identical with this control for practice.

Response times. The median values of correlation coefficients were r = .33 (r_c = .50) for the Target Detection paradigm, r = .34 (r_c = .51) for the Visual Judgment paradigm, and r = .39 (r_c = .56) for the Numeric Judgement paradigm.

Accuracy. The median values of correlation coefficients were r<.0001 (r_c = .001*)* for the Target Detection paradigm, r=.25 (r_c = .40) for the Visual Judgment paradigm, and r=.09 (r_c = .17) for the Numeric Judgment paradigm.

Exploratory analysis of the n–2 repetition cost: Ordering of paradigms. There was a possibility that the poor across-paradigm reliability was a result of a reduction in the n-2 repetition cost as participants progressed through the experimental sessions[6]; due to the counterbalancing, this could mask reliability effects if not controlled.

In order to address this potential issue, we re-categorised the n–2 repetition cost for reach participant as a function of "paradigm order" (i.e., "1st paradigm encountered"; "2nd paradigm encountered"; "3rd paradigm encountered"). We then conducted a one-way ANOVA with the n–2 repetition cost as a dependent variable and the *paradigm order* as the independent variable. This analysis revealed that there was no effect of order on the n–2 repetition cost, $F(2,142) = 0.22$, $p > .8$, $\eta_g^2 = .002$; the 1st paradigm encountered had a mean n–2 repetition cost of 88ms (*SE*=13); the 2nd paradigm encountered had a mean n–2 repetition cost of 95ms (*SE*=13); and the 3rd paradigm encountered had a mean n–2 repetition cost of

[5] We are grateful to Cai Longman for suggesting this possibility.
[6] We are grateful to Cai Longman for suggesting this possibility.

99ms (*SE*=13). Thus, there is no evidence to support a reduction in n–2 repetition cost as participants proceeded through the experiment.

General Discussion

The main aim of the current study was to examine the reliability of the n–2 repetition cost, a promising measure of individual differences in inhibitory control (e.g. Whitmer & Banich, 2007). Consistent with the existing literature, we found large and statistically-significant n–2 repetition costs in all three task switching paradigms used in our study; however, the n–2 repetition cost was found to not be reliable in the split-half reliability analysis, and was not predicted by individual differences hypothesised to be associated with inhibition (i.e., RRS and processing speed).

The current study's findings confirmed the results from Pettigrew and Martin's (2015) study. These authors used the n–2 repetition cost as a component of a battery of tests used to investigate the extent to which individual variation in working memory capacity and interference resolution relates to task switching performance. As part of their study, they assessed the reliability of all of their methods using split-half reliability, and found the n–2 repetition cost to be below the criterion for reliability. Assessing the reliability of this measure was not the primary aim of their study. Our study therefore provides the first systematic examination of the reliability of the n–2 repetition cost using three typical task switching paradigms. Importantly, as each paradigm in the current study had very different task demands, we are more confident that our findings can be generalised.

Our findings present a challenge to researchers wishing to use the n–2 repetition cost for individual differences research, and opens the question as to the cognitive processes "captured" by the n–2 repetition cost; that is, whether the n–2 repetition cost is a measure of a single process (i.e., cognitive inhibition). If a given measure taps into a single process rather

than a number of different processes, that measure would be expected to be reliable (Streiner, 2003). Grange and Juvina (2015) found considerable variation in the n–2 repetition cost of nine subjects in a practice study. Whilst not the focus of their investigation, Grange and Juvina modelled this individual variation by varying a parameter in their computational model that controlled inhibitory input; the model was able to reproduce this individual variation. However, the current findings suggest this individual variation is not reliable, making it harder to interpret. It is possible that as-yet unidentified factors (including individual difference predictors) not controlled for in the current study could have affected the reliability of the n–2 repetition cost.

N–2 Repetition Costs Between Paradigms

An additional feature of using more than one task switching paradigm is that we can assess the degree of correlation of n–2 repetition cost across all paradigms. We found that—whilst the group-level n–2 repetition cost did not vary significantly across different paradigms—individual n–2 repetition costs did not correlate across paradigms. This is converging evidence as to the lack of reliability in the measure: That is, if the n–2 repetition cost is a reliable and accurate measure of an individual's inhibitory control, then a participant with a large n–2 repetition cost in one paradigm should show similar n–2 repetition cost in a different paradigm. Therefore, the finding that the n–2 repetition cost only partially correlated between the three paradigms not only adds to the evidence that this effect is not reliable, but could be taken as evidence for the n–2 repetition cost not being as pure a measure of cognitive inhibition as currently believed. Future research should explore potential reasons for the apparent lack of reliability.

Individual Differences and the N–2 Repetition Cost

It is an under-reported finding that typically the n–2 repetition cost varies considerably between participants (see Figure 2; see also Grange & Juvina, 2015). This individual variation has become of interest to some researchers. The current study aimed to account for known individual differences associated with the n–2 repetition cost, and control for these differences in the reliability analysis; these factors were found not to be associated with the n–2 repetition cost.

Although not the primary focus of our research, our findings did not replicate the negative correlation between depressive rumination and the n–2 repetition cost reported by Whitmer and Banich (2007). Processing speed also did not predict the magnitude of the n–2 repetition cost. The reason for our failure to replicate the negative correlation between depressive rumination and the n–2 repetition cost (Whitmer & Banich, 2007) is uncertain. Chen, Feng, Wang, Su, and Zhang (2016) also found no relationship between the n–2 repetition cost and the level of ruminative tendencies, suggesting more work is required in this area to clarify the true relation between rumination and n–2 repetition costs. Whilst it is of course plausible our findings and those of Chen et al. reflect a genuine failure to replicate Whitmer and Banich's results (e.g., Open Science Collaboration, 2015), we believe methodological differences are a more likely candidate. For example, Whitmer and Banich's Experiment 1 had a sample of forty-three participants selected from the upper and lower 10% of RRS scorers from 776 respondents. Thus, their participants were more "extreme" on the RRS scale than our participants, which could explain our discrepant findings. However, their Experiment 2 had fifty-four participants that were not pre-selected in this manner, and still found a negative correlation. We also note that they statistically controlled for the switch cost which we did not measure in our study; therefore, the difference in statistical control could also explain the discrepancy.

On the Choice of Reliability Criterion

One reviewer raised the possibility that our chosen criterion for acceptable reliability of the n–2 repetition cost— a corrected reliability coefficient greater than .7 —was perhaps too stringent. We are sympathetic to this view because this is the criterion tailored for psychometric tests; experimental cognitive tasks such as the task switching paradigm are perhaps not optimised for individual difference measures. As such, it is perhaps unrealistic to expect this paradigm to reach high levels of reliability.

However, we would like to emphasise that whilst we interpret our data as suggesting the n–2 repetition cost has low reliability, this does not preclude readers reaching different conclusions. Our bootstrapping data provide readers with the distribution of reliability coefficients, and we would advise the reader to use their own judgment on how to interpret these coefficients in terms of reliability. There will likely be variation in these judgements. It is worth mentioning that the level of internal reliability of the n–2 repetition cost obtained from our study has been interpreted before as "moderate" (Leue, Klein, & Lange, 2013), "poor"-to-"good" (Condon et al., 2012), and "quite low" (Pettigrew & Martin, 2015). None of these are necessarily "correct".

Low Reliability Due to Homogenous Sample?

There is a possibility that the homogenous sample used in the current study might have reduced the differential validity by reducing the variance of the n–2 repetition cost, in turn reducing our ability to measure reliability accurately[7]. The current study aimed to obtain data from a type of sample that is comparable to the studies reported in the literature. It was not the focus of this study to investigate how more/ less homogenous sample influences the differential validity and if it affects the variance of the n–2 repetition cost, but this remains an

[7] We are grateful to an anonymous review for suggesting we discuss this.

interesting question for future research. However, based on observations from the current and previous studies, even in a sample as homogenous as university students, there is considerable variance of the n–2 repetition cost (e.g. Grange & Juvina, 2015; see also Figure 4 in the current paper).

N–2 Repetition Cost & Inhibition

The current study has suggested that the n–2 repetition cost has low reliability. Also, the inter-correlations of the n–2 repetition cost between the three task switching paradigms were low. These data are consistent with the notion that the n–2 repetition cost is not a stable measure of inhibitory control. However, these data are also to be expected if the n–2 repetition cost is not a "pure" measure of inhibition; that is, if the n–2 repetition cost arises as a consequence of a mixture of factors—one of which being inhibition—then it is little wonder the reliability is low.

There is some evidence that the n–2 repetition cost is not a pure measure of inhibition. We have recently provided evidence that interference during *episodic retrieval* can affect the magnitude of the n–2 repetition cost (Grange, Kowalczyk, & O'Loughlin, 2015; see also Mayr, 2002; Neill, 1997). This evidence comes from a close replication of Mayr (2002). In his study, Mayr suggested that the n–2 repetition cost could be influenced by episodic retrieval in the following way. When a given task is performed, an episodic trace of the trial's parameters (such as the cue, the stimulus characteristics, and the response made) is stored in memory. When this task is cued again (i.e., task A in an ABA sequence), retrieval of the most recent episodic trace of this task (from n–2) occurs. If parameters of the current trial differ from that of the retrieved episode (e.g., a different response is required), interference occurs between the current trial parameters and the parameters contained within the retrieved trace. This can result in an n–2 repetition cost in the absence of any inhibitory mechanism.

Mayr (2002) developed a paradigm that controlled whether trial parameters matched or mismatched across an ABA sequence; he found that the n–2 repetition cost was not significantly modulated by episodic retrieval. In our close replication (Grange et al., 2015)—with a larger sample size—we found a strong reduction in the n–2 repetition cost (almost half) when controlling for episodic retrieval. That is, we found evidence that if episodic retrieval is not taken into consideration, the estimate of the n–2 repetition cost can be inaccurate.

Related to the current study, one possibility is that the low reliability of the n–2 repetition cost is due to it being a contaminated measure (inhibition plus episodic retrieval). That is, we cannot be certain how much of the variance seen in the n–2 repetition cost in our data is due to episodic retrieval effects. It is likely that not controlling for episodic retrieval affected the reliability of the n–2 repetition cost as well as the inter-correlations between the three paradigms. There is already some evidence (Gade, Souza, Druey, & Oberauer, 2016) that an analogous effect to the n–2 repetition cost present in declarative working memory tasks—the n–2 list repetition cost—seems to be modulated by episodic retrieval effects. With the mentioned studies in mind, our ongoing research is investigating whether reliability of the n–2 repetition cost is influenced by episodic retrieval effects.

Conclusion

Our data suggest the n–2 repetition cost is not reliable as a measure of individual differences in inhibitory control. Individual differences such as tendency to ruminate and processing speed did not predict the n–2 repetition cost. Taken together, these results show that the n–2 repetition cost as a measure of individual differences in inhibitory control should be interpreted with caution.

References

Bestgen, Y., & Dupont, V. (2000). Is negative priming a reliable measure for studying individual differences in inhibition? *Current Psychology of Cognition*, *19*, 287–305.

Brown, L. a, Brockmole, J. R., Gow, A. J., & Deary, I. J. (2012). Processing speed and visuospatial executive function predict visual working memory ability in older adults. *Experimental Aging Research*, *38*(1), 1–19. http://doi.org/10.1080/0361073X.2012.636722

Champely, S. (2009). pwr: Basic functions for power analysis. R package version 1.1.1. Retrieved from http://CRAN.R-project.org/package=pwr.

Chen, X., Feng, Z., Wang, T., Su, H., & Zhang, L. (2016). Internal switching and backward inhibition in depression and rumination. *Psychiatry Research*, *243*, 342–348. http://doi.org/S0165-1781(15)30272-9 [pii]

Congdon, E., Mumford, J. a, Cohen, J. R., Galvan, A., Canli, T., & Poldrack, R. a. (2012). Measurement and reliability of response inhibition. *Frontiers in Psychology*, *3*(February), 37. http://doi.org/10.3389/fpsyg.2012.00037

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. http://doi.org/10.1007/BF02310555

Dreher, J. C., Kohn, P. D., & Berman, K. (2001). Neural basis of backward inhibition during task switching. *NeuroImage*, *13*(6), S311–S311.

Drost, E. A. (2011). Validity and Reliability in Social Science Research. *Education Research and Perspectives*, *38*(1), 105–123.

Fales, C. L., Vanek, Z. F., & Knowlton, B. J. (2006). Backward inhibition in Parkinson's disease. *Neuropsychologia*, *44*(7), 1041–1049. http://doi.org/10.1016/j.neuropsychologia.2005.11.002

Foti, F., Sdoia, S., Menghini, D., Mandolesi, L., Vicari, S., Ferlazzo, F., & Petrosini, L. (2015). Are the deficits in navigational abilities present in the Williams syndrome related to deficits in the backward inhibition? *Frontiers in Psychology*, *6*(March), 1–10. http://doi.org/10.3389/fpsyg.2015.00287

Gade, M., & Koch, I. (2008). Dissociating cue-related and task-related processes in task inhibition: evidence from using a 2:1 cue-to-task mapping. *Canadian Journal of Experimental Psychology*, *62*(1), 51–55. http://doi.org/10.1037/1196-1961.62.1.51

Gade, M., Schuch, S., Druey, M. D., & Koch, I. (2014). Inhibitory Control in Task Switching. In J. A. Grange & G. Houghton (Eds.), *Task Switching and Cognitive Control* (pp. 137–159). Publisher: Oxford University Press.

Goschke, T. (2000). Intentional reconfiguration and involuntary persistence in task set switching. In M. S & J. Driver (Eds.), *Control of cognitive processes (attention & performance series Volume XVIII)* (pp. 331–355). Cambridge, Massachuset: The MIT Press. http://doi.org/10.2337/db11-0571

Grange, J. A. (2010). *Control of cognitive processes in task switching (PhD Thesis)*. UK: Bangor University. http://doi.org/10.1017/CBO9781107415324.004

Grange, J. A., & Houghton, G. (2014). *Task Switching and Cognitive Control*. (J. A. Grange & G. Houghton, Eds.). Oxford Univesity Press.

Grange, J. A., & Juvina, I. (2015). The effect of practice on n-2 repetition costs in set switching. *Acta Psychologica*, *154*, 14–25. http://doi.org/10.1016/j.actpsy.2014.11.003

Grange, J. A., Kowalczyk, A. W., & O'Loughlin, R. (2015). The Effect of Episodic Retrieval on Inhibition in Task Switching. Retrieved from http://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=2482865

Greenberg, J., Reiner, K., & Meiran, N. (2013). "Off with the old": Mindfulness practice improves backward inhibition. *Frontiers in Psychology*, *3*, 1–9. http://doi.org/10.3389/fpsyg.2012.00618

Houghton, G., Pritchard, R., & Grange, J. (2009). The role of cue-target translation in backward inhibition of attentional set. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(2), 466–76. http://doi.org/10.1037/a0014648

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching--a review. *Psychological Bulletin*, *136*(5), 849–874. http://doi.org/10.1037/a0019842

Koch, I., Gade, M., Schuch, S., & Philipp, A. M. (2010). The role of inhibition in task switching: a review. *Psychonomic Bulletin & Review*, *17*(1), 1–14. http://doi.org/10.3758/PBR.17.1.1

Lawo, V., Philipp, A. M., Schuch, S., & Koch, I. (2012). The Role of Task Preparation and Task Inhibition in Age-Related Task-Switching Deficits. *Psychology and Aging*, *27*(4), 1130–1137. http://doi.org/10.1037/a0027455

Leue, A., Klein, C., Lange, S., & Beauducel, A. (2013). Inter-individual and intra-individual variability of the N2 component: On reliability and signal-to-noise ratio. *Brain and Cognition*, *83*(1), 61–71. http://doi.org/10.1016/j.bandc.2013.06.009

Mayr, U. (2001). Age differences in the selection of mental sets: the role of inhibition, stimulus ambiguity, and response-set overlap. *Psychology and Aging*, *16*(1), 96–109. http://doi.org/10.1037/0882-7974.16.1.96

Mayr, U. (2007). Inhibition of Task Sets. In D. S. Gorfein & C. M. MacLeod (Eds.), *Inhibition in Cognition*. APA Books: Washington DC.

Mayr, U., Diedrichsen, J., Ivry, R., & Keele, S. W. (2006). Dissociating task-set selection from task-set inhibition in the prefrontal cortex. *Journal of Cognitive Neuroscience*, *18*(1), 14–21. http://doi.org/10.1162/089892906775250085

Mayr, U., & Keele, S. (2000). Changing internal constraints on action: the role of backward inhibition. *Journal of Experimental Psychology: General*, *129*(1), 4–26. http://doi.org/10.10371/0096-3445.129.1.4

Moritz, S., Hübner, M., & Kluwe, R. (2004). Task Switching and Backward Inhibition in Obsessive-Compulsive Disorder. *Journal of Clinical and Experimental …*, *26*(5), 677–683.

Nolen-Hoeksema, S., & Morrow, J. (1991). A prospective study of depression and posttraumatic stress symptoms after a natural disaster: the 1989 Loma Prieta Earthquake. *Journal of Personality and Social Psychology*, *61*(1), 115–121. http://doi.org/10.1037/0022-3514.61.1.115

Onwuegbuzie, A., & Daniel, L. (2002). A Framework for Reporting and Interpreting Internal Consistency Reliability Estimates. *Measurement and Evaluation in Counseling and Development*, *35*, 89–103.

Pettigrew, C., & Martin, R. C. (2015). The Role of Working Memory Capacity and Interference Resolution Mechanisms in Task Switching. *The Quarterly Journal of Experimental Psychology*, *November*, 1–43. http://doi.org/10.1080/17470218.2015.1121282

Philipp, A. M., & Koch, I. (2006). Task inhibition and task repetition in task switching. *European Journal of Cognitive Psychology*, *18*(4), 624–639. http://doi.org/10.1080/09541440500423269

Philipp, A. M., & Koch, I. (2009). Inhibition in language switching: what is inhibited when switching between languages in naming tasks? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(5), 1187–1195. http://doi.org/10.1037/a0016376

Picardi, C. A., & Masick, K. D. (2013). Research Methods: Designing and Conducting Research with a Real-world Focus. *SAGE Publications*, (April), 43–53.

Prior, A. (2012). Too much of a good thing: Stronger bilingual inhibition leads to larger lag-2 task repetition costs. *Cognition*, *125*(1), 1–12. http://doi.org/10.1016/j.cognition.2012.06.019

Revelle, W., & Condon, D. M. (2014). Reliability. In P. Irwing, T. Booth, & D. Hughes (Eds.), *Handbook of Psychometric Testing*. Wiley-Blackwell.

Schuch, S., & Koch, I. (2003). The role of response selection for inhibition of task sets in task shifting. *Journal of Experimental Psychology: Human Perception and performanceHuman Perception and Performance*, *29*(1), 92–105. http://doi.org/10.1037/0096-1523.29.1.92

Stawski, R. S., Sliwinski, M. J., & Hofer, S. M. (2013). Between-person and within-person associations among processing speed, attention switching, and working memory in younger and older adults. *Experimental Aging Research*, *39*(2), 194–214. http://doi.org/10.1080/0361073X.2013.761556

Strauss, G. P., Allen, D. N., Jorgensen, M. L., & Cramer, S. L. (2005). Test-retest reliability of standard and emotional stroop tasks: an investigation of color-word and picture-word versions. *Assessment*, *12*(3), 330–7. http://doi.org/10.1177/1073191105276375

Streiner, D. L. (2003). Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment*, *80*(1), 99–103.

http://doi.org/10.1207/S15327752JPA8001_18

Thanoi, W., & Klainin-Yobas, P. (2015). Assessing rumination response style among undergraduate nursing students: A construct validation study. *Nurse Education Today*, *35*(5), 641–646. http://doi.org/10.1016/j.nedt.2015.01.001

Treynor, W., Gonzalez, R., & Nolen-Hoeksema, S. (2003). Rumination reconsidered: A psychometric analysis. *Cognitive Therapy and Research*, *27*(3), 247–259. http://doi.org/10.1023/A:1023910315561

Unsworth, N., Heitz, R. P., & Engle, R. W. (2005). An automated version of the operation span task. *Behaviour Research Methods*, *37*(3), 498–505.

van der Elst, W., van Boxtel, M. P. J., van Breukelen, G. J. P., & Jolles, J. (2006). The Letter Digit Substitution Test: normative data for 1,858 healthy participants aged 24-81 from the Maastricht Aging Study (MAAS): influence of age, education, and sex. *Journal of Clinical and Experimental Neuropsychology*, *28*(6), 998–1009. http://doi.org/10.1080/13803390591004428

Vandierendonck, A., Liefooghe, B., & Verbruggen, F. (2010). Task switching: interplay of reconfiguration and interference control. *Psychological Bulletin*, *136*(4), 601–626. http://doi.org/10.1037/a0019791

Whitmer, A. J., & Banich, M. T. (2007). Inhibition versus switching deficits in different forms of rumination: Research article. *Psychological Science*, *18*, 546–553. http://doi.org/10.1111/j.1467-9280.2007.01936.x

Whitmer, A. J., & Banich, M. T. (2012). Brain activity related to the ability to inhibit previous task sets: an fMRI study. *Cognitive, Affective, & Behavioral Neuroscience*, 661–670. http://doi.org/10.3758/s13415-012-0118-6

Yiu-kwan, E. (2008). *Backward Inhibition in Pathological Gamblers*. Hong Kong. Retrieved from http://hdl.handle.net/10722/55143

Figure Captions

Figure 1. Schematic overview of the trial structure for each of the task switching paradigms. Note the images are not to scale.

Figure 2. Density functions of the response time (RT) n–2 repetition cost distribution for each paradigm, calculated as RT (ABA) – RT (CBA).

Figure 3. Density functions of the accuracy n–2 repetition cost distribution for each paradigm, calculated as % Accuracy (ABA) – % Accuracy (CBA).

Figure 4. Reliability checks for the n–2 repetition cost for response times. The plots show violin plots of the (uncorrected) bootstrapped split-half reliability estimates (correlation coefficients, *r*) each paradigm. The horizontal dashed line represents the criteria for reliability ($r \approx .5385$; Picardi & Masick, 2013; Revelle & Condon, 2014).

Figure 5. Reliability checks for the n–2 repetition cost for accuracy. The plots show violin plots of the (uncorrected) bootstrapped split-half reliability estimates (correlation coefficients, *r*) each paradigm. The horizontal dashed line represents the criteria for reliability ($r \approx .5385$; Picardi & Masick, 2013; Revelle & Condon, 2014).

Table 1. Mean response times (in milliseconds) and accuracy (in percent) for ABA and CBA sequences for each paradigm. Standard errors are shown in parentheses. The n–2 repetition cost is calculated as ABA – CBA for both response time and accuracy. Note that for accuracy a negative n–2 repetition cost reflects poorer accuracy on ABA trials.

	Response Times		Accuracy		N–2 repetition cost	
Paradigm	ABA	CBA	ABA	CBA	RT	Accuracy
Target Detection	1014	923	97.20	98.12	91	-.92
	(27)	(28)	(.24)	(.18)		
Visual Judgment	1280	1175	96.94	97.27	105	-.33
	(41)	(38)	(.24)	(.25)		
Numeric Judgment	1323	1237	96.27	96.27	86	0
	(45)	(42)	(.28)	(.31)		

Table 2. Correlation matrix for the mean response times from the three task switching tests, individual differences, age, sex, and the break (gap) between the two experimental sessions.

	Numeric	Target	Visual	RRS	Processing	Age	Sex	Gap
Numeric	—							
Target	.64**	—						
Visual	.75**	.68**	—					
RRS	.02	.10	.19	—				
Processing	-.49**	-.63**	-.49**	-.08	—			
Age	.35*	.25*	.30*	-.02	-.21	—		
Sex	.06	.14	.03	-.21	-.21	.11	—	
Gap	.02	-.03	-.05	.12	-.02	-.13	.19	—

Note: ***p*<.0018 (Bonferroni-corrected criterion for significance). **p*<.05

Table 3. Correlation matrix for the mean overall accuracy from the three task switching tests, individual differences, age, sex, and the break (gap) between the two experimental sessions.

	Numeric	Target	Visual	RRS	Processing	Age	Sex	Gap
Numeric	-							
Target	.42**	-						
Visual	.50**	.63**	-					
RRS	.14	.27*	.22	-				
Processing	.05	.08	-.07	-.08	-			
Age	-.04	.07	.01	-.02	-.21	-		
Sex	-.08	-.13	.02	-.21	-.21	.11	-	
Gap	-.17	.12	.01	.12	-.02	-.13	.19	-

Note: ***p*<.0018 (Bonferroni-corrected criterion for significance). **p*<.05

Table 4. Summary of multiple regression analyses for the n–2 repetition cost as the dependent variable and the processing speed and RRS scores ad independent variables.

		t	***p***	**β**	***F***	***df***	***p***	**adj.R^2**
	Predictor							
Target	RRS	-.01	.99	-.03				
	Processing	.14	.90	.11				
	Overall model				.01	2,68	.99	-.03
Visual	RRS	1.30	.38	2.06				
	Processing	-.74	.46	-.68				
	Overall model				.72	2,68	.49	-.00
Numeric	RRS	1.18	.24	3.91				
	Processing	-1.13	.26	-1.45				
	Overall model				1.46	2,68	.24	.01

Table 5. Correlation matrix for the n–2 repetition cost in response times from the three task switching tests, individual differences, age, sex, and the break (gap) between the two experimental sessions.

	Numeric	Target	Visual	RRS	Processing	Age	Sex	Gap
Numeric	—							
Target	.25*	—						
Visual	.30*	.07	—					
RRS	.15	.01	.12	—				
Processing	-.15	.02	-.10	-.08	—			
Age	.14	-.10	-.02	-.02	-.21	—		
Sex	.06	-.04	-.05	-.21	-.21	.11	—	
Gap	.09	.07	.17	.12	-.02	-.13	.19	—

Note: ***p*<.0018 (Bonferroni-corrected criterion for significance). **p*<.05
With equal number of trials in each paradigm there was only one significant correlation: the n–2 repetition cost between the Visual Judgment and the Numeric Judgment Paradigm, $r(71) = .29$, $p = .01$; this does not survive the criterion for Bonferroni-corrected significance.

Table 6. Correlation matrix for the n–2 repetition cost in accuracy from the three task switching tests, individual differences, age, sex, and the break (gap) between the two experimental sessions.

	Numeric	Target	Visual	RRS	Processing	Age	Sex	Gap
Numeric	—							
Target	.06	—						
Visual	.01	-.15	—					
RRS	.16	.00	-.03	—				
Processing	.02	-.06	.00	-.08	—			
Age	-.09	-.07	.11	-.02	-.21	—		
Sex	-.18	-.08	-.10	-.21	-.21	.11	—	
Gap	-.08	-.07	-.10	.12	-.02	-.13	.19	—

Note: None of the correlations were significant, all *ps > 0.1*.