

Alternatives to switch-cost scoring in the task-switching paradigm: Their reliability and increased validity

Meredith M. Hughes · Jared A. Linck · Anita R. Bowles ·
Joel T. Koeth · Michael F. Bunting

Published online: 20 December 2013
© Psychonomic Society, Inc. 2013

Abstract In the task-switching paradigm, the latency switch-cost score—the difference in mean reaction time between switch and nonswitch trials—is the traditional measure of task-switching ability. However, this score does not reflect accuracy, where switch costs may also emerge. In two experiments that varied in response deadlines (unlimited vs. limited time), we evaluated the measurement properties of two traditional switch-cost scoring methods (the latency switch-cost score and the accuracy switch-cost score) and three alternatives (a rate residual score, a bin score, and an inverse efficiency score). Scores from the rate residual, bin score, and inverse efficiency methods had comparable reliability for latency switch-cost scores without response deadlines but were more reliable than latency switch-cost scores when higher error rates were induced with a response deadline. All three alternative scoring methods appropriately accounted for differences in accuracy switch costs when higher error rates were induced, whereas pure latency switch-cost scores did not. Critically, only the rate residual and bin score methods were more valid indicators of task-switching ability; they demonstrated stronger relationships with performance on an independent measure of executive functioning (the antisaccade analogue task), and they allowed the detection of larger effect sizes when examining within-task congruency effects. All of the three alternative scoring methods provide researchers with a better measure of task-switching ability than do traditional scoring methods, because they each

simultaneously account for latency and accuracy costs. Overall, the three alternative scoring methods were all superior to the traditional latency switch-cost scoring method, but the strongest methods were the rate residual and bin score methods.

Keywords Task switching · Switch cost · Measurement · Scoring method · Reliability · Validity

Task-switching ability is one of many executive control abilities, the set of general mechanisms that regulate the performance of other cognitive processes (Miyake et al., 2000), but the measurement properties of task-switching scores are often ignored. Variations in task-switching ability can appear through switch costs in latency and accuracy, but traditional latency switch-cost scores fail to provide a single, comprehensive score that incorporates both latency and accuracy. This is problematic when conducting group comparisons (e.g., Prior & MacWhinney, 2010) and even more so when examining individual differences (e.g., Friedman & Miyake, 2004; Friedman et al., 2006; Miyake et al., 2000).

In the task-switching paradigm, participants perform two simple tasks in serial alternation, switching between them following either a predictable rule or an external cue (for reviews, see Logan, 2003, Monsell, 2003, & Vandierendonck, Liefoghe, & Verbruggen, 2010). The stimuli afford responses to both tasks, so the participant relies on the rule or cue to know which task to perform on any given trial. Trials directly following a change in the task to be performed (*switch trials*) are typically slower and/or less accurate than trials that do not directly follow a change in task (*nonswitch trials*). Figure 1 contains examples of trial sequences in a task-switching task during two pure conditions and the mixed condition, when participants switch between tasks.

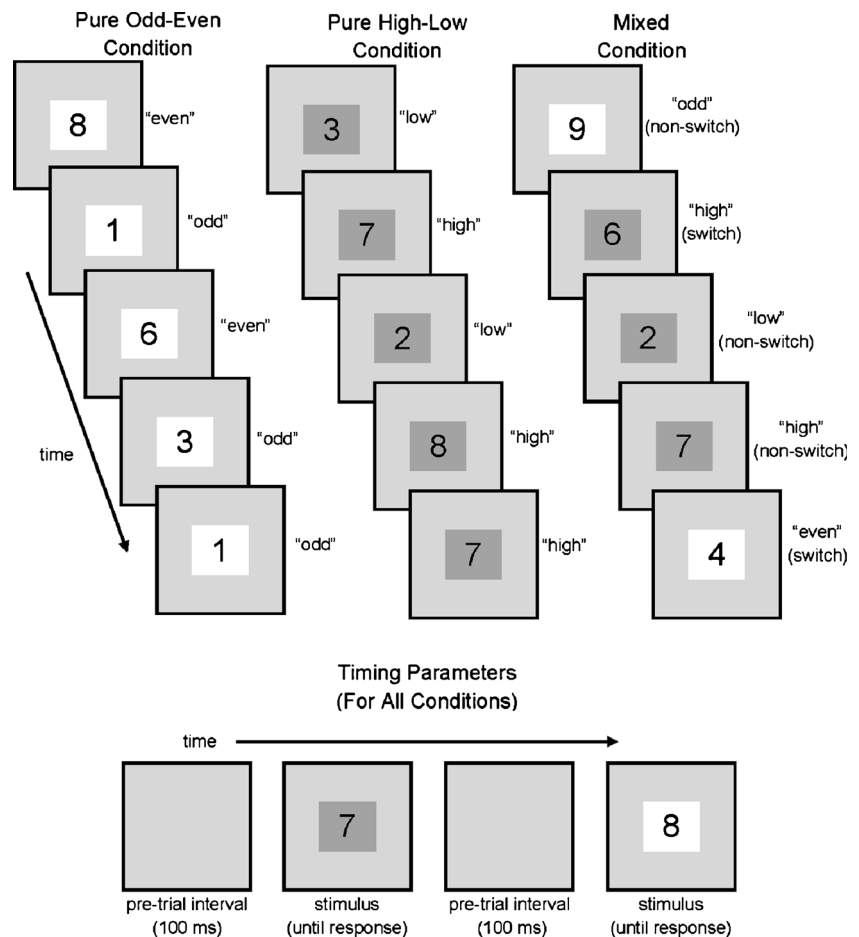
Typically, switching between tasks has costs to speed and/or accuracy. A latency *switch-cost* score reflects the additional time required to switch from one task to the other (i.e., the difference in reaction time [RT] between switch and

Electronic supplementary material The online version of this article (doi:10.3758/s13428-013-0411-5) contains supplementary material, which is available to authorized users.

M. M. Hughes (✉) · J. A. Linck · A. R. Bowles · M. F. Bunting
Center for Advanced Study of Language, University of Maryland,
7005 52nd Avenue, College Park, MD 20742, USA
e-mail: mhughes@casl.umd.edu

J. T. Koeth
Second Language Acquisition Program, University of Maryland,
7005 52nd Avenue, College Park, MD 20742, USA

Fig. 1 Schematic of task-switching task during pure/mixed conditions in Experiment 1. In pure conditions, participants (1) judge whether the digit is *odd* or *even* or (2) whether the digit is *lower* or *higher than five*. In the mixed condition, participants alternate between tasks every three trials. Background color corresponds to task: White means *odd* versus *even*, dark gray means *lower* versus *higher*. Correct responses are in quotations, and for mixed condition, trial type is in parentheses. The bottom portion shows timing parameters.



nonswitch trials). Likewise, an accuracy switch-cost score is the difference in error rates. The goal of the present study was to compare the latency and accuracy of switch-cost scoring methods with those of three alternative methods—a rate residual score, a bin score, and an inverse efficiency score—on the basis of reliability and validity. The validities of the scoring methods were compared on (1) their ability to capture switch costs exhibited through changes in both response speed and accuracy, (2) the strength of their relationship to performance on the antisaccade analogue task (e.g., Hallett, 1978), a measure of the related but separate executive function of inhibitory control, and (3) their ability to detect within-task congruency effects.

There is still debate on the exact mechanisms responsible for switch costs (for a review, see Kiesel et al., 2010). Some argue that switch costs reflect a shifting mechanism responsible for the reconfiguration of the current task set, whereas others believe that costs are driven solely by proactive interference occurring within the context of associative retrieval processes (e.g., Koch & Allport, 2006; Monsell, 2003). Indeed, there is a rich literature examining task-switching performance to identify the component processes that support such goal-directed behavior (e.g., De Jong, 2000; Jamadar, Michie, & Karayanidis, 2009; Meiran, 2000; Rogers &

Monsell, 1995; Rubinstein, Meyer, & Evans, 2001). The present study was not designed to adjudicate among differing theoretical perspectives; rather, it demonstrates that the choice of scoring methods can impose serious limitations on study conclusions, thereby impacting theory development and model testing. For example, whereas some researchers have found a significant interaction between task switch and congruency, such that larger switch costs are observed on incongruent trials (Gopher, Armony, & Greenspan, 2000, Experiment 2), several other studies have failed to detect the interaction (Fagot, 1994; Gopher et al., 2000, Experiment 1; Meiran, 1996; Monsell, Azuma, Eimer, Le Pelley, & Strafford, 1998). One potential explanation for the discrepancy of results is that this congruency effect is a true effect but it is not detected by all scoring methods, due to some aspect of a scoring method (e.g., reliability). If this were true, given that these congruency effects have informed theoretical accounts of task switching (Kiesel et al., 2010), then the choice of scoring method has clear implications for the field's ability to advance the state of current theoretical models.

Investigations into the effects of different features of the task, such as the type of cuing and predictability of switches, have helped to identify different processing components (for recent examples, see Meier, Woodward, Rey-Mermet, & Graf,

2009; Saeki & Saito, 2009). Such studies often involve the parametric manipulation of multiple task features, and researchers test for complex higher-order (e.g., three-way) interactions among these factors. Other researchers have used the task-switching paradigm as an individual-difference measure, to examine the relationship of task switching to other variables of interest, such as age, performance on related cognitive tasks, presence of disorders (e.g., schizophrenia), and even the use of energy supplements (for examples, see Einöther, Martens, Rycroft, & De Bruin, 2010; Gupta, Kar, & Srinivasan, 2009; Jamadar, Michie, & Karayanidis, 2010; Poljac et al., 2010), or to examine the cognitive consequences of being bilingual (Prior & Gollan, 2011; Prior & MacWhinney, 2010).

For all types of studies, reliable measurement of the variation in individuals' task-switching ability has implications for drawing inferences and conclusions. Researchers taking an experimental approach are often interested in examining complex interactions between different factors (e.g., task manipulations) in order to better understand the mental processes supporting task performance. Although analyses typically aggregate across individuals and focus on group-level effects, unreliable measurement of task performance introduces irrelevant noise that could obfuscate the targeted higher-order interactions. In contrast, the primary goal of individual-difference studies is typically to measure a person's underlying ability by use of a single score (e.g., switch cost, mix cost) or by estimating a latent variable based on multiple scores or "indicators" (for examples of a latent variable analysis with measures of executive control, see Friedman & Miyake, 2004; Miyake et al., 2000). A latent variable analysis has the benefit of partialing out measurement error, thereby providing an estimate of the true underlying construct. But latent variable analysis requires large sample sizes (hundreds of participants) and the collection of multiple indicators per construct, both of which may not be possible for most researchers due to logistical, financial, or other constraints. Therefore, most researchers often rely on a single score to measure a construct. In this case, lower reliability may lead to erroneous measurement of the participants' abilities, thereby making it more difficult to detect a true relationship between task switching and other measures or outcome variable. Therefore, when using the task-switching paradigm, to allow valid inferences to be drawn, it is important to consider the reliability and validity of the resulting scores.

Reliability

Reliability concerns the ability of a score to consistently measure relative differences in ability (over different samples of items, test forms, testing occasions, etc.). Lower levels of reliability are problematic because scores with low reliability contain a comparatively high proportion of error variance,

which constrains the researcher's confidence that the variation in observed scores represents variation in the intended construct. Error variance also negatively affects the likelihood of observing significant relationships with other variables by attenuating or obscuring such relationships. With decreased confidence in the representativeness of scores and higher amounts of error variance in scores, low reliability affects theory and model development, a common problem in cognitive control research (Rabbitt, 1997).

Although standards for reliability can vary greatly (by purpose, domain, or context), Nunnally and Bernstein (1994) recommend a reliability criterion of .70 for experimental and developmental purposes. However, not all researchers investigate or report measures of reliability. Others report reliabilities that do not consistently meet Nunnally and Bernstein's criteria for experimental and developmental purposes. For example, Salthouse, Fristoe, McGurthy, and Hambrick (1998) reported reliabilities of .71, .61, and .46 for the latency switch-cost measures in their study, and Miyake et al. (2000) reported reliabilities of .91 and .59 for the latency switch-cost measures in their study. In these cases, the reported levels of reliability vary widely, even within a single study.

Validity

Validity concerns the extent to which variation in scores reflects individual differences in an underlying ability of interest. Low validity is problematic because scores with low validity contain higher proportions of construct-irrelevant variance. The present study focused on the threat to the validity of latency switch-cost scores from ignoring individual differences in accuracy switch costs, or vice versa. Although latency and accuracy are known to be highly related (Pachella, 1974), participants may engage in particular response strategies that prioritize accuracy over speed (i.e., speed-accuracy trade-offs; e.g., Samavatyan & Leth-Steensen, 2009). Using latency switch-cost scores as the sole indicator of individual differences in task-switching ability essentially ignores half of the data—accuracy. Because switching between tasks can result in costs to both speed and accuracy (Monsell, 2003), measuring costs to RT, but not accuracy, captures only a portion of the available information about a participant's performance.

The prioritization of speed to the point of sacrificing accuracy is particularly problematic for latency switch-cost scores. Wickelgren (1977) found that when experiments are manipulated to emphasize speed (e.g., using response deadlines or special instructions), more erroneous responses can be expected. This reduction in accuracy may more strongly impact switch trials, due to the increased difficulty of these trials and the presumed need to engage additional processes. Even

if not experimentally induced, participants may strategically decide to emphasize the speed of their performance despite negative effects on accuracy. Such depreciations in accuracy, however, are not reflected in traditional RT measures, like the latency switch-cost score. Moreover, since data from incorrect trials are typically excluded from RT estimates, speed–accuracy trade-offs may also negatively impact the reliability of RT estimates by reducing the number of observations contributing to the latency score.

Researchers often conduct separate analyses on RT and accuracy costs, then examine whether the patterns of results are parallel across the two outcomes (e.g., Rubin & Meiran, 2005). Unless the specific intent of the researcher is to compare costs exhibited in latency with costs exhibited in accuracy, the strategy of separate analyses represents a potential threat to the power of the researcher to detect relationships and interactions. Separate analyses essentially ignore one reflection of cost (either latency or accuracy) at a time. If some individuals tend to exhibit costs in one variable more than in the other or even switch between experiencing more cost to accuracy or to speed of performance, the effect that the researcher is attempting to detect will be weakened and, therefore, more difficult to detect, when speed and accuracy are analyzed separately. One possible extension to separate ANOVA analyses could be to perform a MANOVA, treating RT and accuracy as dependent variables simultaneously. A MANOVA, however, requires a number of assumptions (e.g., linear dependency between dependent variables, multivariate normal distributions) that RT and accuracy data are not likely to meet and has no clear application in individual-difference research contexts.

In sum, utilizing a score with low reliability and/or validity has the potential to harm a researcher's theory and model building by leading to the assignment of noisy and/or biased scores to participants. Here, we consider alternative scores aimed at capturing both latency and accuracy costs in a single score.

Scoring methods

Given these concerns about task-switching scores' reliability and validity, the focus of the present research was to improve the measurement of individual differences in task-switching ability. Three alternative scoring methods for the task were compared with standard latency-only and accuracy-only switch-cost scores. The new scoring methods were selected because they simultaneously incorporate information about the speed *and* the accuracy of responses. The present analyses compare the traditional latency and accuracy switch-cost scores with three alternative scoring methods—a rate residual score, a bin score (see below for details), and an inverse efficiency score. The methods were compared in terms of their

reliability, their ability to simultaneously measure costs to latency and accuracy, their ability to detect within-task congruency effects, and their relationship with a separate measure of inhibitory control—the antisaccade analogue task. Previous research has found that task switching and inhibition are related but separable components of executive functioning (e.g., Miyake et al., 2000), so inhibitory control performance should show a relationship to task switching performance. The scoring methods were compared in two experimental contexts, which investigated the impact of the new scoring methods across different levels of error rates.

Latency and accuracy switch-cost scores

The traditional latency switch-cost score is calculated using RTs from within mixed-condition block(s). A participant's average RT on accurate nonswitch trials is subtracted from his or her average RT on accurate switch trials. In calculating both averages, RTs from inaccurate trials are excluded. The accuracy switch-cost score is calculated by subtracting the proportion of correct responses on switch trials from that on nonswitch trials.

The primary drawback of the latency switch-cost score is that, by focusing only on the speed of responses and excluding the inaccurate trials from the mean RT calculation, changes in accuracy are not reflected in a participant's latency switch-cost score. Conversely, the drawback of the accuracy switch-cost score is that changes in latency are not reflected in a participant's accuracy switch-cost score.¹

Rate residual scoring method

The rate residual score is modeled on the scoring method used with a test of available long-term memory access (Was & Woltz, 2007; Woltz & Was, 2006). The rate residual score incorporates speed and accuracy information by computing the residualized difference between the *rates of correct responses per second* for switch and nonswitch trials. By examining correct response rates, rather than RTs on correct responses, this measurement of task performance takes into account both response latencies and response accuracy. For example, if a participant is responding with a similar speed on switch and nonswitch trials but has lower accuracy on switch trials, the average RTs may be similar, but the number of correct responses for switch trials (the numerator of their switch trial rate) will be lower, thus incorporating costs to accuracy into the score. Note that, for this score, more negative scores indicate larger costs, scores near zero indicate

¹ It should be noted that some researchers examine switch costs in latency and accuracy separately (Monsell, 2003) or look for speed–accuracy trade-offs (e.g., Gopher et al., 2000; Samavatyan & Leth-Steensen, 2009). However, these approaches fail to incorporate both aspects of performance into a single score or into a single analysis.

average, or typical, costs, and more positive scores indicate smaller costs relative to the sample of participants (for more detail, see the [Method](#) section).

Was and Woltz (2007) cited two reasons for using residualized differences over simple difference scores. First, residualized differences are typically more reliable than simple difference scores, especially when the two difference components become more highly correlated (Linn & Slinde, 1977). Since the present goal is to improve the measurement qualities of task-switching scores, this is also an advantage for the rate residual method. Second, Was and Woltz believed simple difference scores to be problematic for measuring individual differences in available long-term memory access because the magnitude of priming depends partly on overall speed and accuracy (there is less room for the facilitation effect they were measuring when an individual was already fast and accurate). Similarly, the rate residual score proposed for task switching takes overall baseline performance into account, such that the same raw costs are not necessarily considered to reflect the same level of task-switching ability given differences in baseline performance.

Bin scoring method

In the bin score (modeled on procedures proposed in Fidelman & Kenyon, 2009), each switch trial RT is subtracted from the participant's average RT for all nonswitch trials. The RT differences between each single switch trial RT and the average nonswitch trial RT for the sample of participants are rank-ordered and placed into ten bins. The smallest differences (i.e., the smallest switch costs) are placed in the “best” bin, and the greatest differences (i.e., largest switch costs) are placed in the “worst” bin. The bins are labeled numerically with the integers 1–10 to create an ordinal variable where the “better” bins have values closer to one and the “worse” bins have values closer to ten. The RT differences for each participant are recoded to the value of the bin they belonged to and summed to compute the participant's bin score. In this way, on the basis of RTs alone, participants whose switch trial RTs tend to be only slightly longer than their nonswitch trial RTs would have a smaller bin score.

To incorporate information on accuracy into this score, inaccurate responses are penalized by automatically placing those trials in a “bad” bin. Thus, participants who make many errors on switch trials would receive higher bin scores, indicating worse performance. We note that the assignment of inaccurate responses to a bad, or the worst, bin is arbitrary. However, any choice of penalty would be arbitrary to some degree. The goal of this approach was to add an additional cost for errors, and it is unclear whether RTs from inaccurate trials merit interpretation similar to those on correct trials. That is, simply adjusting the bin by a set amount (e.g., +3) for making an error does not carry any strong theoretical justification over

our approach, which implies that an inaccurate response is twice as undesirable as a latency difference exhibiting the greatest degree of switch cost. The bin score considered here penalized errors by assigning them a bin value of 20 (i.e., double the value of the “worst” correct trial RT bin), thereby imposing a stronger penalty for committing an error.²

Inverse efficiency scoring method

The inverse efficiency scoring method was proposed by Townsend and Ashby (1978, 1983) as a way to integrate speed and accuracy for analysis. Inverse efficiency scores incorporate information about speed and RT by dividing RTs by 1 minus the percentage of errors (i.e., percentage of correct responses). Bruyer and Brysbaert's (2011) review of inverse efficiency scores showed mixed results, such that inverse efficiency scores had advantages over RT analysis in some situations (e.g., when error rates are low and there is a strong, positive linear relationship between speed and accuracy), but not in others (e.g., when error rates are high and there is not a strong, positive linear relationship between speed and accuracy).

Experiment 1

Participants in Experiment 1 completed a task-switching task and an antisaccade analogue task designed to measure the inhibition component of executive control. In the task-switching task, participants were instructed to respond as quickly and accurately as possible, in contrast to Experiment 2, which included a response deadline. The three alternative scoring methods were compared with the traditional latency and accuracy switch cost scores with respect to reliability and criterion validity (i.e., the correlation with antisaccade performance).

Method

Participants

Participants were studying foreign languages as part of their training for the U.S. military. In total, 1,902 participants (423 females, 1,304 males, and 175 participants with unreported gender; $M_{\text{age}} = 24.17$ years, $SD = 4.68$) produced usable data. One hundred and two participants, 5.1 % of the sample, were excluded from analyses for exhibiting both low overall RT and

² We also examined an alternative bin score whereby errors were assigned to the worst RT bin—that is, assigned a bin value of 10. However, the results consistently favored the double-penalty binning approach; therefore, we only report the results for the double-penalty bin score.

accuracy in the task-switching task (less than 25th percentile of mean overall RT and less than .80 overall accuracy) because that level of performance was considered to reflect low motivation to complete the task as directed.

Because the study was being conducted at the request of their institution and while participants were on duty, participants were given a half-day's release from their normal schedules to participate in the study. Care was taken to ensure that participants knew that their participation was optional, and the procedures used were approved by a university human subjects' Institutional Review Board (IRB) and a Department of Defense IRB.

Procedure

The task-switching and antisaccade analogue tasks were administered as part of a larger test battery (see Bunting et al., 2010). All participants completed one session, during which they completed a background history questionnaire and a number of computerized tasks. The computerized tasks measured aspects of working memory, rote memory, primability, implicit induction, and auditory perceptual acuity. The testing session was broken into three subsections, each taking approximately 45 min. The test order was fixed for all participants with the task-switching and antisaccade analogue tasks administered as the seventh and fourth tasks, respectively. A subset of participants also completed a separate session during which additional cognitive ability tests and scales of personality and motivational traits were administered.

Materials

All tasks were implemented in E-Prime (version 2.0, Psychology Software Tools), and responses were recorded with a serial-port response box (Psychology Software Tools).

Task-switching task Stimuli consisted of the digits 1–9, excluding 5, and afforded performance on two tasks: the odd–even (OE) task and the high–low (HL) task. The OE task required participants to classify the stimulus number as either “odd” or “even.” For all OE trials, the target digit appeared centered in a white background box. The HL task required participants to classify the stimulus number as either “lower than 5” or “higher than 5.” For all HL trials, the target digit appeared centered in a gray background box. During the mixed condition (shown in Fig. 1, including timing parameters), participants were instructed to switch between tasks every three trials. In addition, task switches were externally cued, with either a white or a gray background square serving as a visual reminder of the to-be-

completed task. Response box templates were created for the task so that button one had a label of Odd/Low and button two had a label of Even/High above the corresponding buttons. The assignment of button labels was held constant across participants, since the goal of these studies was to examine individual differences in processing, rather than responses averaged over the group.

Participants first received instructions for the OE task and then completed 8 practice trials with feedback, followed by 16 practice trials without feedback. Following this practice, they completed 32 baseline OE test trials (a pure-condition block). Participants then followed the same procedure for the HL task—instructions, 8 practice trials with feedback, 16 practice trials without feedback, and 32 test trials. Then participants completed 24 mixed-condition block practice trials with feedback and 24 practice trials without feedback. They then completed two mixed-condition test blocks of 72 test trials (without feedback) each, with a short break between the two blocks. Thus, across the two test blocks for the mixed condition, participants completed a total of 46 switch trials and 98 nonswitch trials.

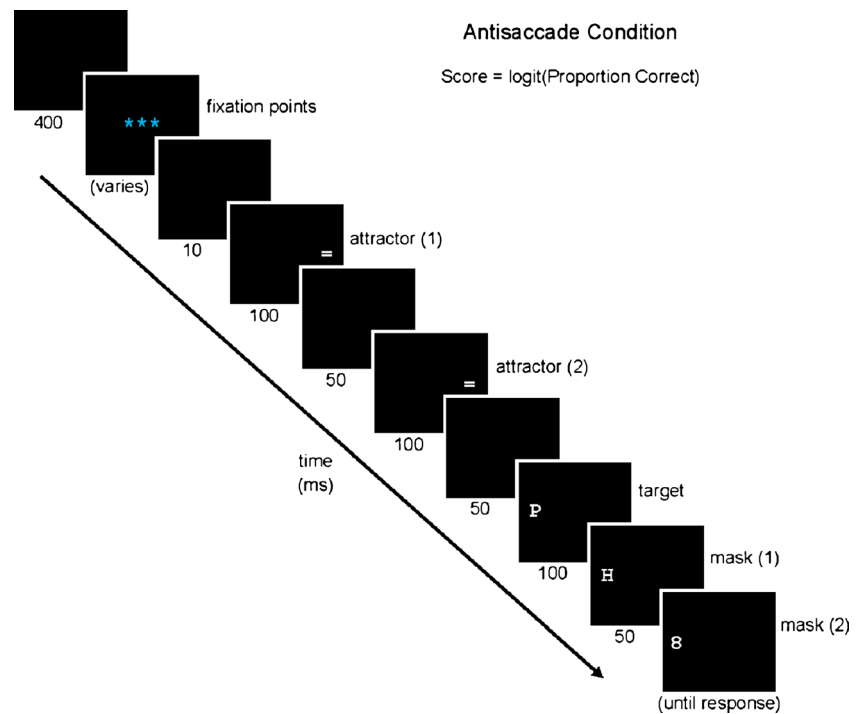
Each participant completed the trials in the same, fixed order. Test trials for the mixed-condition blocks were pseudorandomized with the following constraints:

- 1) the same stimulus digit could not repeat across trials, with at least two intervening trials before the digit could be repeated;
- 2) the same target response button could repeat no more than three times in a row;
- 3) no more than five trials in a row were *button-response congruent*, such that the correct response would require pressing the same button regardless of the current task.

Antisaccade analogue task The antisaccade analogue task is a manual response analogue of the traditional antisaccade eye movement task (Hallett, 1978) and was included in the analyses as a measure of inhibition. Task set switching and inhibition have been found to be related but separate components of executive function (e.g., Miyake et al., 2000). Therefore, we expected performance on this task to be related to task-switching performance.

The methodology and procedure for this task were adapted from Unsworth, Schrock, and Engle (2004). In the antisaccade analogue task, participants must resist their prepotent tendency to orient attention toward a distractor stimulus and, instead, direct visual attention to the opposite side of a computer screen to identify a target stimulus. On a given trial, a distractor signal (“=”) is flashed twice on one side of the screen, followed immediately by the brief presentation of a backward-masked target letter (P, B, or R; see Fig. 2). The participant must indicate which target letter was presented by

Fig. 2 Schematic of the antisaccade analogue task during the critical antisaccade block. A cue is flashed on the opposite side of the screen from where the backward masked target letter will subsequently appear. Thus, participants must overcome their prepotent response to look at the cue and, instead, fixate the target letter prior to the backward mask to correctly identify the letter and make the buttonpress response. The duration of the fixation point screen was 200, 600, 1,000, 1,400, or 1,800 ms.



pressing one of three response buttons.³ Previous research using this task has found that individuals with better executive functioning are more able to suppress the prepotent response to look at the distractor signal and, thus, are more accurate at identifying the target stimuli. Performance on the antisaccade analogue task was measured as a logit transformation of the proportion of accurate trials.

Scoring

The traditional and alternative scoring methods are described below.

Traditional latency switch-cost score Prior to calculating the switch and nonswitch average RTs, RT data were cleaned to remove inaccurate and outlying trials. Those trials following an inaccurate trial were also excluded. To identify outlying trials, the standard deviation (*SD*) and mean RTs were calculated for each participant within each of the four conditions, crossing switch condition (switch vs. nonswitch) with response task (HL vs. OE). RTs below 200 ms were replaced

with the participant's mean for that condition. Long outlying RTs were defined as being more than 3 *SD*s above the participant's conditional mean and were replaced with that participant's cutoff. After the RT outliers were replaced with their appropriate values, the final RT averages for switch and nonswitch conditions (and collapsing HL and OE conditions) were calculated.⁴

Latency switch costs were computed by subtracting the average nonswitch RT from the average switch RT for each participant, such that larger (more positive) values indicated a larger cost to switching.

Traditional accuracy switch-cost score Accuracy switch costs were based on the accuracy of responses in the mixed-condition block. A participant's accuracy switch cost was computed by subtracting the proportion of correct responses on switch trials from the proportion of correct responses on nonswitch trials. As with the latency switch-cost score, a larger (more positive) value indicated a larger cost to switching.

Rate residual score To calculate the rate residual score, the RT data were cleaned as described above. The test block was then divided into subsets of trials, excluding the initial 3 nonswitch

³ Note that the antisaccade analogue task is a three-alternative forced choice (3AFC) task, whereas the task-switching task was a 2AFC task. Although this introduces a methodological difference between the tasks, we believe that this is a benefit to our research goals: design-specific variance (or "common method bias") can inflate or deflate estimates of an effect (for an analytic derivation with linear and interaction effects in regression, see Siemsen, Roth, & Oliveira, 2010). By reducing design-specific shared variance, we can better estimate the true relationship between the task-switching scoring methods and performance on the antisaccade task.

⁴ Since there is not strong agreement on best-practices for trial-cleaning procedures, we followed the advice of a reviewer who suggested exploring whether minor alterations to the trial-cleaning procedure would affect our conclusions. The results in Tables 3, 4, 5, 6 and 7 were re-created following two variations in trial-cleaning: excluding RT outliers and including recovery trials (i.e., trials following an error). We found that neither change would have affected the present conclusions.

trials of each mixed-condition test block.⁵ Assigning the trials into eight subsets resulted in most subsets containing 6 switch trials and 12 nonswitch trials. The exceptions were the fourth and eighth subsets, which contained 5 switch and 10 nonswitch trials.

Within each subset, the rate of correct responses per second was calculated for the nonswitch trials and switch trials by dividing the number of correct responses per trial type by the time taken to make all of the responses, whether accurate or inaccurate (summing the RTs for that trial type). As an example, the formula for the rate of correct responses for nonswitch trials, $R_{\text{Nonswitch}}$, in one subset of trials is given in Eq. 1, where $c_{\text{Nonswitch}}$ is the number of correct responses to nonswitch trials in this subset and $RT_{\text{Nonswitch}}$ is the response time, regardless of accuracy, for each nonswitch trial in this subset:

$$R_{\text{Nonswitch}} = \frac{c_{\text{Nonswitch}}}{\sum RT_{\text{Nonswitch}}} \quad (1)$$

For each trial subset (e.g., subset 1), the switch trial rates were regressed on nonswitch trial rates. Equation 2 is an example of the regression model for predicting a participant's rate for switch trials (R_{Switch}) with the rate for nonswitch trials ($R_{\text{Nonswitch}}$) within a subset of trials:

$$\widehat{R}_{\text{Switch}} = \beta_0 + \beta_1(R_{\text{Nonswitch}}) + \varepsilon. \quad (2)$$

Although participants were expected to have a lower rate for the switch trials, the residuals—the difference between observed and predicted switch trial rates—indicated which participants exhibited relatively greater or lesser “costs” in their switch rates, as compared with the rest of the sample (see Eq. 3 for an example of a single subset). Participants who were affected (i.e., slowed down) on switch trials more than expected earned negative residuals,⁶ whereas those who slowed down on switch trials less than expected earned positive residuals. That is, a larger switch cost is indicated by more negative residuals, using this measure. Since the regression was performed on each of the eight subsets of trials (indexed by j in Eq. 4), a total of eight residuals were computed for each participant. The residuals from each of the J regressions were averaged to find each participant's final rate residual score, RR , as seen in Eq. 4:

⁵ Note that the number of subsets the data is divided into is somewhat arbitrary. Differences in task design may change the ideal number of subsets into which data should be divided, but the goal is to support the internal consistency of the scores by achieving a balance between the number of residuals to average and the number of trials per person that each regression is based on.

⁶ Recall that larger rates indicate greater speed, in contrast to RTs, where smaller RTs indicate greater speed.

$$e = R_{\text{Switch}} - \widehat{R}_{\text{Switch}} \quad (3)$$

$$RR = \frac{\sum_{j=1}^8 e_j}{8}. \quad (4)$$

Bin score The bin score compared each trial's switch trial RT with the individual's average nonswitch RT. First, the RT data were cleaned as described above, and the participants' average nonswitch RTs were calculated. The nonswitch means were then subtracted from each switch trial RT, resulting in an RT difference for each switch trial for each person. The differences from all of the participants' accurate trials were then binned according to their size, as shown in the top portion of Fig. 3. The first decile of the differences between the nonswitch RT means and each switch trial were assigned to bin 1, the second decile were assigned to bin 2, and so on. This resulted in ten equi-percentile bins, with responses in bin 1 exhibiting the least cost from a switch and bin 10 exhibiting the most cost. Finally, all inaccurate trials, which had been excluded during the binning process, were assigned a bin value of 20. By giving errors an artificial assignment to bin 20, the bin score doubles the penalty for inaccurate responses, relative to the worst correct trial values, and thus increases the negative impact of errors to the participants' bin scores. To compute a participant's final bin score, the bin labels assigned to all of his or her RT differences were summed. This process is illustrated for a hypothetical participant in the bottom portion of Fig. 3.

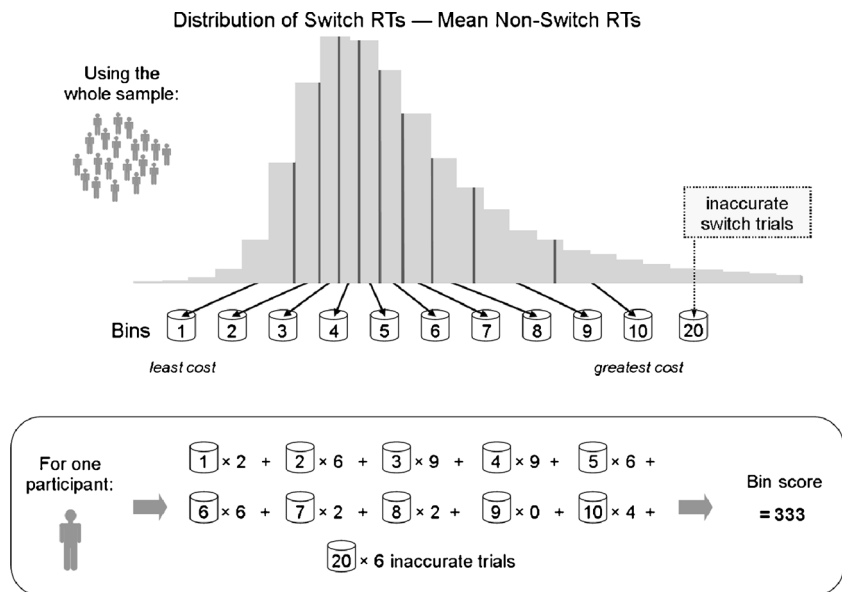
Inverse efficiency score To calculate the inverse efficiency scores, the RT data were cleaned as described above. Then, separately for switch and nonswitch trials, inverse efficiency was computed by dividing the mean RT by the mean accuracy. The final inverse efficiency score was the difference between the switch and nonswitch inverse efficiency subscores (i.e., switch inverse efficiency – nonswitch inverse efficiency). As with the latency and accuracy switch-cost scores, a larger (more positive) value indicated a larger cost to switching.

Data analysis

Reliability Internal consistency coefficients reflect how consistently participants or examinees perform over a set of items (Crocker & Algina, 2006). The internal consistencies of the rate residual and bin scores were estimated with Cronbach's coefficient alpha (Cronbach, 1951). Alpha was computed over each participant's residual terms (8 items) for the rate residual score and the participant's bin values (46 items) for the bin scores.

As was described above, the latency and accuracy switch-cost scores and inverse efficiency scores are difference scores; the reliability of a difference score is calculated using Eq. 5:

Fig. 3 Schematic of the bin scoring procedure. The top portion of the figure shows the distribution of switch reaction times (RTs)—average nonswitch RTs for the sample of participants. The smallest differences are placed in bin 1, the bin associated with the least switch cost. The largest differences are placed in bin 10, the bin associated with the greatest switch cost. Inaccurate responses to switch trials are placed in bin 20. In the bottom portion, the score for 1 participant is shown as a sum of each bin label times the number of trials the participant had assigned to that bin.



$$\rho_{DD'} = \frac{\rho_{XX'}\sigma_X^2 + \rho_{YY'}\sigma_Y^2 - 2\rho_{XY'}\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2 - 2\rho_{XY'}\sigma_X\sigma_Y}, \tag{5}$$

where $\rho_{DD'}$ is the reliability of the difference score ($X - Y$), $\rho_{XX'}$ and $\rho_{YY'}$ are the reliabilities of the components X and Y , σ_X and σ_Y are their standard deviations, and $\rho_{XY'}$ is their correlation (Zimmerman & Williams, 1982). To find the reliabilities of the accuracy switch-cost components ($\rho_{XX'}$ and $\rho_{YY'}$), alpha was computed across response accuracy for the 46 and 98 switch and nonswitch trials. To find the reliabilities of the RT means and inverse efficiency scores for each condition ($\rho_{XX'}$ and $\rho_{YY'}$), odd–even split-half⁷ estimates of internal consistency were used in place of coefficient alpha because the exclusion of inaccurate trials resulted in each participant having slightly different amounts and patterns of missing data. There were 72 trials in the even-numbered and odd-numbered test halves.

Assessing the impact of alternative scoring methods Two methods were used to assess the ability of different scoring methods to simultaneously incorporate both latency and accuracy costs into one score. First, the relationship between the latency switch-cost scores and the rate residual, bin, and inverse efficiency scores were visually compared between participants who experienced the lowest and highest levels of accuracy switch costs. Second, the amount of variance in rate residual, bin, and inverse efficiency scores explained by latency switch-cost scores was compared with the amount of variance explained by both latency and accuracy switch costs.

⁷ Correlation between scores based on odd-numbered and even-number trials, corrected according to the Spearman–Brown prediction formula (Brown, 1910; Spearman, 1910).

Relationship with an external criterion Further analyses compared the task-switching scores’ relationships with scores on the antisaccade analogue task—a measure of the related executive control function of inhibition. The relationships of the traditional and proposed task-switching scores with the antisaccade analogue score were examined using Spearman’s rank correlation coefficient, ρ , to account for any curvilinear relationships. Additionally, the amount of variance in the antisaccade analogue scores explained by the latency switch costs and accuracy switch costs (both separately and in combination) was compared with the amount of variance explained by the rate residual, bin, and inverse efficiency scores. To the extent that the alternate scoring methods reduce measurement error by incorporating both latency and accuracy into a single score, we should find a stronger relationship between the alternate scores and the antisaccade analogue scores.

Assessing a switch and congruency interaction The final analysis comparing the scoring methods assessed the extent to which congruency effects were detected by each scoring method. The interaction between switch and congruency conditions was assessed by applying the scoring methods to the button-response congruent trials and to the button-response incongruent trials separately. The differences between scores for the congruent and incongruent conditions reflected the interaction between switch and congruency conditions. Incorporating latency and accuracy information into a single score, as in the proposed task-switching scores, was hypothesized to increase the power to detect such interactions. We expected to find larger switch and congruency interaction effects when assessing the effect size of the differences between congruent task-switching scores and incongruent task-switching scores, as indicated by Cohen’s d for correlated means.

Results

Descriptive statistics for the task-switching scores and antisaccade analogue scores are given in Table 1. Descriptive statistics for overall accuracy and mean latencies during the pure-condition test blocks are given in Table 2.

Reliability

The internal consistency estimates obtained for the latency switch-cost scores, accuracy switch-cost scores, rate residual scores, bin scores, and inverse efficiency scores are given in Table 3. The accuracy switch cost had by far the poorest internal consistency ($\alpha = .226$). Note that the latency switch-cost, rate residual, bin, and inverse efficiency scores meet the suggested minimum of .70 for research and development contexts and exhibited comparable levels of internal consistency.

Assessing the impact of alternative scoring methods

To demonstrate the impact of the alternative scores, we examined the relationships of latency switch-cost scores to the three alternative scores for participants exhibiting varying levels of accuracy switch costs.

Figure 4 plots participants' rate residual scores against their latency switch-cost scores, for participants in the first and fourth quartiles⁸ of the distribution of accuracy switch-cost scores. Participants in the first quartile, who exhibited the smallest accuracy switch costs, tended to have larger rate residual scores (i.e., smaller switch costs) than did participants in the fourth quartile, who exhibited the greatest accuracy switch costs. This tendency indicates that, for participants with comparable latency switch-cost scores, the rate residual score appropriately penalizes participants with greater accuracy switch costs by assigning them lower rate residual scores. Similarly, Fig. 5 plots participants' bin scores against their latency switch-cost scores. Recall that, contrary to rate residual scores (where better task-switching ability is reflected by larger scores), better task switching is indicated by smaller bin scores. Participants in the first quartile tended to have smaller bin scores than did participants in the fourth quartile, indicating that the bin score incorporated a penalty for participants who exhibited greater accuracy switch costs, given comparable latency switch costs. Finally, Fig. 6 plots participants' inverse efficiency switch-cost scores against their latency switch-cost scores. Participants in the first quartile tended to have smaller inverse efficiency scores than did participants in the fourth quartile, indicating that the inverse efficiency scores incorporated a penalty for participants with a given level of

latency switch cost who exhibited greater accuracy switch costs.

The regression models showed whether differences in accuracy switch costs explained differences in rate residual, bin, and inverse efficiency scores above and beyond the contribution of latency switch costs. Table 4 provides the unstandardized regression weights (B), their standard errors (SE), and the change in the proportion of variance explained (R^2) for the regression models. In step 1 of the regressions, the models included an intercept term and latency switch-cost scores. Accuracy switch-cost scores were added to the model in step 2. When predicting rate residual scores, adding accuracy switch costs as a predictor explained 3.3 % ($p < .001$) more of the variance in rate residual scores than did latency switch-costs alone. For bin scores, adding accuracy switch costs explained 7.8 % ($p < .001$) more of the variance in bin scores. For inverse efficiency switch costs, latency switch costs alone accounted for 92.8 % of variance in the inverse efficiency scores, but adding the accuracy switch-cost scores explained an additional 4.4 % of unique variance. The large and significant increases in R^2 for rate residual, bin, and inverse efficiency scores demonstrate that, although latency switch costs explain a large proportion of the variation in these scores, additional unique variance can be explained by differences in accuracy switch costs. The necessity of both latency and accuracy switch costs for explaining variation in the rate residual, bin, and inverse efficiency scores demonstrates that the three alternative scores capture participants' differences in both latency and accuracy switch costs with a single measure. Note, however, that bin scores had the largest amount of unique variance explained by accuracy switch costs, suggesting that bin scores may provide the best method for modeling latency and accuracy differences in a single score.

Relationships with antisaccade analogue task performance

Table 5 provides the correlations between the task-switching scores and the antisaccade analogue scores. The correlation matrix shows that the relationships between the latency switch-cost scores and the proposed task-switching scoring methods were moderately-to-fairly strong. It is important to note that the alternative scores maintained a strong relationship with the traditional scoring method, providing convergent validity evidence and indicating that they were able to capture much of the individual variation measured by the latency switch-cost scores. The accuracy switch cost's smaller correlations with the alternative scores are likely due to the extremely low reliability of the accuracy switch-cost score.

Of greater importance were the differences in correlations with the external criterion: antisaccade analogue scores. Table 5 shows that, of the five task-switching scoring methods explored, the traditional latency and accuracy switch-cost scores had the weakest correlations with antisaccade analogue scores.

⁸ Participants in the middle quartiles were not plotted to highlight the differences of the extreme groups and minimize over plotting.

Table 1 Descriptive statistics

Score	<i>N</i>	<i>M</i>	<i>SD</i>	Min	Max	Skew	Kurtosis
Experiment 1							
RT switch cost	1,902	418.85	173.08	5.87	1,204.05	0.52	0.09
Accuracy switch cost	1,902	0.02	0.04	−0.10	0.20	0.78	1.48
Rate residual	1,902	0.00	1.27	−3.50	5.55	0.38	0.37
Bin	1,902	289.87	66.58	105	513	0.21	−0.17
IES switch cost	1,902	448.04	197.03	7.02	1,288.55	0.42	0.24
Overall accuracy	1,902	0.96	0.03	0.81	1.00	−1.46	2.69
Overall mean RT	1,902	825.04	129.77	479.32	1,288.55	0.42	0.24
Antisaccade analogue	1,784	0.68	0.90	−1.70	3.37	0.25	−0.29
Experiment 2							
RT switch cost	46	62.66	33.60	3.95	167.67	0.95	1.02
Accuracy switch cost	46	0.17	0.10	−0.04	0.45	0.27	0.29
Rate residual	46	0.00	1.57	−4.85	3.27	−0.44	0.19
Bin	46	1,432.26	172.05	1137	1958	0.65	0.43
IES switch cost	46	323.92	231.09	28.73	1,392.98	2.25	7.68
Overall accuracy	46	0.67	0.03	0.59	0.72	−0.29	−0.34
Overall mean RT	46	429.37	74.58	330.73	711.37	1.34	2.78
Antisaccade analogue	44	0.70	0.93	−0.85	2.40	0.03	−1.21

Note. IES = inverse efficiency score. Experiment 1 skew *SE* = 0.06, kurtosis *SE* = 0.11; Experiment 2 skew *SE* = 0.35, kurtosis *SE* = 0.69.

Although all three alternative scoring methods had stronger relationships with the antisaccade analogue scores, as compared with the latency and accuracy switch-cost scores, relationships of similar magnitude were found for the rate residual scores and the bin scores ($r_s = .280$ and $-.253$, respectively), and these correlations were larger than that found with the inverse efficiency switch-cost score ($r = -.208$).

The results of regressing antisaccade analogue scores on the switch-cost, rate residual, bin, and inverse efficiency scores, provided in Table 6, also demonstrate the strength of the relationship between performance on the antisaccade analogue task and the task-switching task as scored using different methods. The rate residual and bin scores performed the best, since they explained more variance in the antisaccade analogue scores ($R^2_s = .079$ and $.064$ for Models 2 and 3 in Table 6, respectively⁹) than did the latency switch-cost scores, the accuracy switch-cost scores, both the latency and accuracy switch-cost scores, and both scores plus an interaction term ($R^2_s = .034$, $.015$, $.045$, and $.047$ for models 1.1–1.4 in Table 6, respectively). Unlike the rate residual and bin scores, the inverse efficiency score (Model 4) failed to outperform Models 1.3 and 1.4, with $R^2 = .043$, suggesting that the inverse efficiency score may not adequately capture accuracy switch costs in addition to latency switch costs.

The greater R^2_s of Models 2 (rate residual) and 3 (bin) indicate that the rate residual and bin scores can explain greater variance in antisaccade analogue scores than is

possible using the four combinations of latency and accuracy switch-cost scores in Models 1.1–1.4. The greater reliability of the rate residual and bin scores and the removal of the systematic bias introduced by ignoring either accuracy or latency costs likely strengthened the observable relationships between the task-switching task and the antisaccade analogue task (i.e., validity was enhanced).

Assessing a switch and congruency interaction

The Cohen's d effect sizes and mean scores for congruent and incongruent trials for each scoring method are provided in Table 7.¹⁰ Although each scoring method revealed an interaction between switch and congruency, the effect size of the interaction was larger for the rate residual and bin scores than for the latency, accuracy, and inverse efficiency switch-cost scores. This indicates that the proposed scoring methods had more power to detect the interaction effect, likely because they combined two sources of information (latency and accuracy costs) into a single metric. Note that the inverse efficiency switch-cost score largely failed to increase the power to detect this interaction effect.

Experiment 2

Experiment 2 explored the question of whether the alternate scoring methods would provide increased reliability

⁹ As noted in the Discussion, the strengths of these relationships are comparable to those found by Miyake et al. (2000), Friedman and Miyake (2004), and Friedman et al. (2006).

¹⁰ The rate residual scores were based on regressions dividing the trials into four, rather than eight, subsets of trials to maintain approximately the same number of trials in the regressions after splitting the trials by congruency. This was also done for the parallel analyses in Experiment 2.

Table 2 Experiment 1 pure test block descriptive statistics

Condition	Score	<i>N</i>	<i>M</i>	<i>SD</i>	Min	Max	Skew	Kurtosis
Odd–Even	Overall accuracy	1,902	0.96	0.06	0.09	1.00	–5.68	53.80
	Overall mean RT	1,902	563.47	83.88	39.41	1,008.41	0.37	3.69
High–Low	Overall accuracy	1,902	0.95	0.08	0.13	1.00	–4.97	30.79
	Overall mean RT	1,902	539.62	86.57	52.56	936.09	0.13	3.39

Note. Skew *SE* = 0.06, kurtosis *SE* = 0.11.

and validity when a task manipulation encouraged participants to commit more errors than in the previous experiment. In Experiment 1, participants were instructed to respond “as quickly and as accurately” as they were able. As a result, the overall error rate was relatively low (.04). Because the alternate scoring methods were proposed to incorporate variation in accuracy and accuracy switch costs, the high level of overall accuracy likely led to an underestimation of the extent to which the proposed scoring methods may improve the measurement of task-switching ability. To address this issue, Experiment 2 manipulated the degree to which participants were encouraged to optimize speed over accuracy. In this experiment, a response deadline was set such that participants needed to respond within a given time window, modeled after the methods of Lien, Ruthruff, Remington, and Johnston (2005). This manipulation was intended to encourage participants to optimize their response speed, likely resulting in decreased response accuracy.

Method

Procedure

The task-switching and antisaccade analogue tasks were administered during a single computerized session. The testing

Table 3 Internal consistency estimates

Score	<i>N</i>	Reliability
Experiment 1		
RT switch cost	1,902	.859
Accuracy switch cost	1,902	.226
Rate residual	1,902	.854
Bin	1,902	.840
IES switch-cost	1,902	.840
Experiment 2		
RT switch cost	46	.484
Accuracy switch cost	46	.470
Rate residual	46	.723
Bin	46	.831
IES switch cost	46	.630

Note. IES = inverse efficiency score.

session took approximately 40 min. The task-switching task was administered first, and the antisaccade analogue task was administered second. Participants were compensated for participating in the study (\$20).

Participants

Participants were students at a local university. In total, 46 participants produced usable data (1 participant, 2.1 % of the sample, was excluded from analyses for low overall RT and accuracy in the task-switching task). The procedures used were approved by a university human subjects' IRB.

Materials

As in the previous experiments, all tasks were implemented in E-Prime (version 2.0, Psychology Software Tools), and responses were recorded with a serial-port response box (Psychology Software Tools).

Deadline task-switching task Stimuli and the response tasks for the deadline version of the task-switching task were as described in Experiment 1. As in Experiment 1, during the mixed condition, participants were instructed to switch between tasks every three trials. In addition, task switches were externally cued, with either a white or a gray background square serving as a visual reminder of the to-be-completed task. Participants were instructed that they would have a deadline before which they needed to respond, described in more detail below. Participants received feedback after every trial throughout the practice and tests blocks, with a response–stimulus wait-time of 100 ms. If the participant responded correctly prior to the response deadline, a yellow smiley face was displayed for 300 ms. If the participant responded incorrectly prior to the response deadline, a red frowning face was displayed for 850 ms, creating a delay that encouraged participants to respond correctly if possible. If the participant did not respond prior to the response deadline, the message, “No response detected. Please respond more quickly,” was presented for 850 ms, again creating a delay to encourage participants to respond as quickly and as accurately as possible.

Response box templates were identical to those in Experiment 1, and the assignment of button labels was held constant across participants.

Fig. 4 Reaction time (RT) switch-cost scores by rate residual scores. Participants with accuracy switch costs in the first quartile (exhibiting the smallest costs) are plotted as circles and are smoothed with a solid LOESS (local regression) line. Participants with accuracy switch costs in the fourth quartile (exhibiting the greatest costs) are plotted as ×'s and are smoothed with a dashed LOESS line. Recall that high levels of task-switching ability are reflected by smaller RT switch-cost scores and larger rate residual scores.

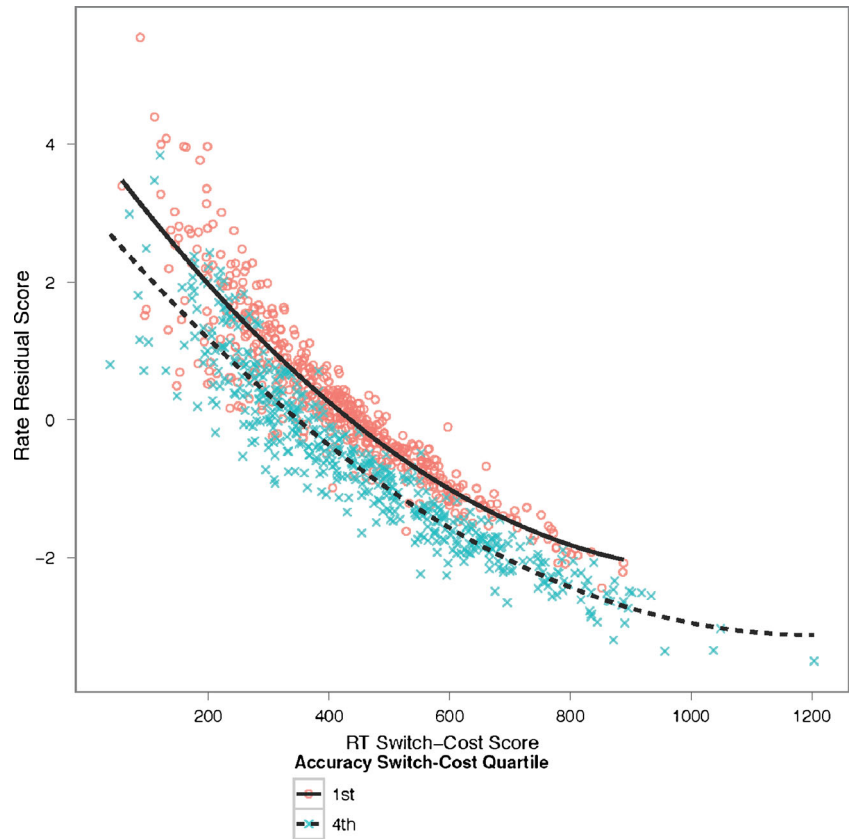


Fig. 5 Reaction time (RT) switch-cost scores by bin scores. Participants with accuracy switch costs in the first quartile (exhibiting the smallest costs) are plotted as circles and are smoothed with a solid LOESS line. Participants with accuracy switch costs in the fourth quartile (exhibiting the greatest costs) are plotted as ×'s and are smoothed with a dashed LOESS line. Recall that high levels of task-switching ability are reflected by smaller RT switch-cost scores and bin scores.

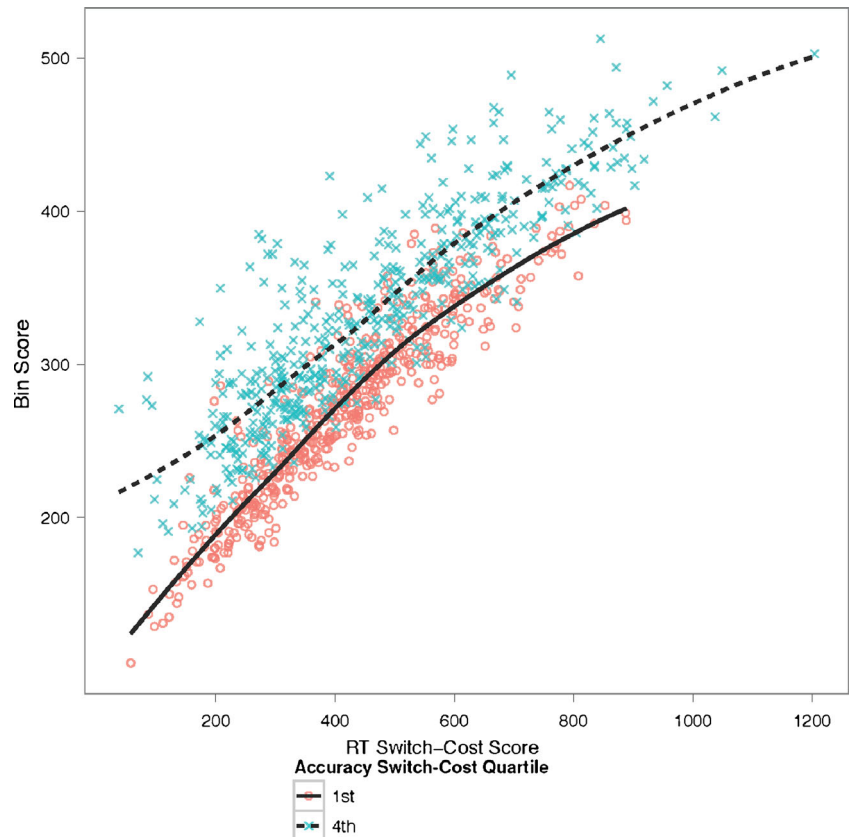
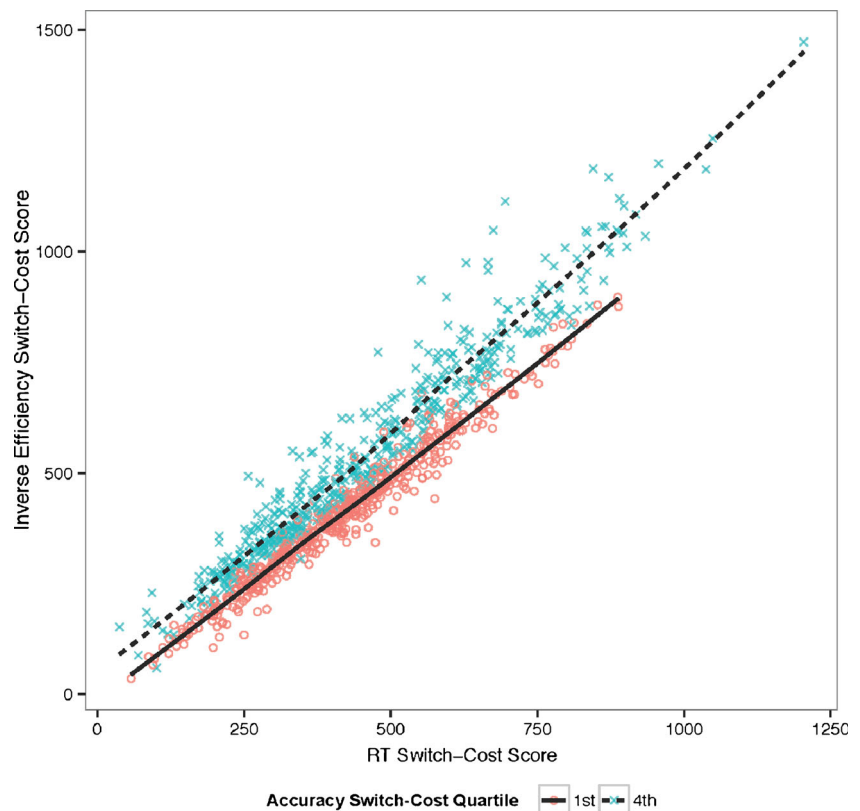


Fig. 6 Reaction time (RT) switch-cost scores by IES switch-cost scores. Participants with accuracy switch costs in the first quartile (exhibiting the smallest costs) are plotted as circles and are smoothed with a solid LOESS line. Participants with accuracy switch costs in the fourth quartile (exhibiting the greatest costs) are plotted as ×'s and are smoothed with a dashed LOESS line. Recall that high levels of task-switching ability are reflected by smaller RT switch-cost scores and IES switch-cost scores.



Participants first received instructions for the OE task and then completed 8 practice trials with feedback, with a fixed response deadline of 3,000 ms. Participants then followed the same procedure for the HL task—instructions and 8 practice trials with feedback and a response deadline of 3,000 ms. Then participants completed 48 mixed-condition block practice trials with feedback. The response deadline decreased steadily over the mixed-condition practice trials from 3,000 to 1,026 ms, decreasing by 42 ms each trial. Participants then completed eight mixed-condition test blocks of 48 test trials each, with a short break between each block. Thus, across the eight test blocks for the mixed condition, participants completed a total of 120 switch trials and 264 nonswitch trials. The response deadline during the mixed-condition test blocks was tailored to each participant on the basis of the accuracy and speed of their responses to nonswitch trials, separately for each condition (OE and HL). The initial response deadline for every participant for each condition was 1,000 ms. Following Lien, Ruthruff, Remington, and Johnston (2005), the deadline was decreased when a participant succeeded on two consecutive nonswitch trials of the condition, and the deadline was increased when a participant failed on a nonswitch trial. The increased deadlines were the product of the current deadline and the

ratio: $(14 + \text{block number}) / (13 + \text{block number})$. The decreased deadlines were computed by dividing the current deadline by the same ratio.

Each participant completed the trials in 8 stimulus lists randomly selected from 20 possible pseudorandomized lists. Test trials for the mixed-condition block stimuli lists were pseudorandomized with the following constraints:

- 1) the same stimulus digit could not repeat across trials, with at least two intervening trials before the digit could be repeated;
- 2) the same target response button could repeat no more than three times in a row;
- 3) no more than 5 trials in a row were *button-response congruent*, such that the correct response would require pressing the same button regardless of the current task.

Antisaccade analogue task The antisaccade analogue task was identical to that used in the previous experiment.

Scoring

The scoring methods were identical to those in the previous experiment, with the exception of the rate residual score due to the different number of trials in

Table 4 Results of regressing rate residual, bin, and IES on switch costs

Predictor	<i>B</i>	(<i>SE</i>)	ΔR^2
Rate Residual			
Step 1			.823***
(Intercept)	2.778	0.032	
RT switch cost	-6.632×10^{-3}	7.060×10^{-5}	
Step 2			.033***
(Intercept)	2.882	0.003	
RT switch cost	-6.501×10^{-3}	6.399×10^{-5}	
Accuracy switch cost	-5.669	0.271	
Total R^2	.856***		
Bin			
Step 1			.783***
(Intercept)	147.310	1.863	
RT switch cost	-.340	0.004	
Step 2			.078***
(Intercept)	143.808	1.498	
RT switch cost	0.330	0.003	
Accuracy switch cost	456.347	14.050	
Total R^2	.860***		
IES Switch Cost			
Step 1			.928***
(Intercept)	-11.348	3.171	
RT switch cost	1.097	0.007	
Step 2			.044***
(Intercept)	-19.160	1.976	
RT switch cost	1.073	0.004	
Accuracy switch cost	1018.000	18.530	
Total R^2	.860***		

Note. $N = 1,902$. IES = inverse efficiency score.

*** $p < .001$

this version of the task. Assigning the trials into eight subsets (by test block) resulted in subsets containing 15 switch trials and 31 nonswitch trials.

Table 5 Correlation between task-switching and antisaccade analogue scores

Score	Latency Switch Cost	Accuracy Switch Cost	Rate Residual	Bin	IES Switch Cost
Experiment 1					
Accuracy switch cost ^a	.098***				
Rate residual ^a	-.907***	-.270***			
Bin ^a	.885***	.364***	-.891***		
IES switch cost ^a	.963***	.303***	-.918***	.934***	
Antisaccade analogue ^b	-.185***	-.121***	.280***	-.253***	-.208***
Experiment 2					
Accuracy switch cost ^c	.211				
Rate residual ^c	-.477***	-.758***			
Bin ^c	.411**	.924***	-.830***		
IES switch cost ^c	.389**	.895***	-.847***	.928***	
Antisaccade analogue ^d	.098	.210	.124	-.060	.069

Note. IES = inverse efficiency score.

^a $N = 1,902$, ^b $N = 1,784$, ^c $N = 46$, ^d $N = 44$.

* $p < .05$

** $p < .01$

*** $p < .001$.

Data analysis

Reliability The internal consistencies of the rate residual and bin scores were estimated with Cronbach's coefficient alpha. Alpha was computed over each participant's residual terms (8 items) for the rate residual score and the participant's bin values (120 items) for the bin score. To find the reliabilities of the accuracy and latency switch-cost components and inverse efficiency components ($\rho_{XX'}$ and $\rho_{YY'}$), a split-half estimate of internal consistency was used in place of coefficient alpha, because the item covariance matrix was not invertible, and the exclusion of inaccurate trials resulted in each participant having slightly different amounts and patterns of missing data. There were 192 trials in the even-numbered and odd-numbered test halves.

Relationship with an external criterion The final analyses comparing the scoring methods examined the relationships of the task-switching scores with scores on the antisaccade analogue task. The relationships of the latency and accuracy switch-cost, rate residual, bin, and inverse efficiency scores with the antisaccade analogue score were examined using Spearman's rank correlation coefficient.

Results

Descriptive statistics for the task-switching scores and antisaccade analogue scores are given in Table 1. The accuracy levels in this task were substantially lower than in Experiment 1 (67 %, vs. 96 %), indicating that the response deadline procedure in Experiment 2 successfully induced a higher error rate.

Reliability

The internal consistency estimates obtained for the latency switch-cost scores, accuracy switch-cost scores, rate residual

Table 6 Results of regressing antisaccade analogue scores on task-switching scores

Predictor	<i>B</i>	(<i>SE</i>)	<i>R</i> ²
Switch-Costs			
Model 1.1: RT switch cost (Intercept)	1.083	0.054	.034***
RT switch cost	-9.631×10^{-4}	1.211×10^{-4}	
Model 1.2: ACC switch cost (Intercept)	0.727	0.023	.015***
Accuracy switch cost	-2.647	0.515	
Model 1.3: RT & ACC switch cost (Intercept)	1.101	0.055	.045***
RT switch cost	-9.100×10^{-4}	1.211×10^{-4}	
Accuracy switch cost	-2.269	0.509	
Model 1.4: Interaction RT & ACC Switch-cost (Intercept)	1.154	0.063	.047***
RT switch cost	-1.035×10^{-3}	1.411×10^{-4}	
Accuracy switch cost	-4.320	1.298	
RT × accuracy switch cost	0.004	0.003	
Rate Residual			
Model 2: Rate residual (Intercept)	0.681	0.020	.079***
Rate residual	.199	0.016	
Bin			
Model 3: Bin (Intercept)	1.674	0.093	.064***
Bin	-3.427×10^{-3}	3.110×10^{-4}	
IES Switch Cost			
Model 4: IES switch cost (Intercept)	1.106	0.052	.043***
IES switch cost	-9.491×10^{-4}	1.057×10^{-4}	

Note. Based on data from Experiment 1, $N = 1,784$. IES = inverse efficiency score.

*** $p < .001$.

Table 7 Switch × congruency interaction effect sizes

Score	Cohen's <i>d</i>	<i>M</i>	
		Congruent	Incongruent
Experiment 1			
RT switch cost	0.114	410.428	426.089
Accuracy switch cost	0.348	0.002	0.030
Rate residual	0.560	0.202	-0.131
Bin ^a	0.558	5.837	6.727
IES switch cost	0.298	420.458	476.773
Experiment 2			
RT switch cost	0.333	57.359	72.995
Accuracy switch cost	0.503	0.147	0.198
Rate residual	1.586	0.695	-0.682
Bin ^a	2.050	10.319	13.539
IES switch cost	0.851	224.512	528.419

Note. IES = inverse efficiency score.

^a Bin scores calculated as the average of bin labels, rather than sum of bin labels, to account for unbalanced number of congruent and incongruent trials.

scores, bin scores, and inverse efficiency scores are given in Table 3. The accuracy switch-cost and latency switch-cost scores had the poorest internal consistency ($\alpha = .470$ and $.484$, respectively). Although neither meet the standards for research or developmental purposes, this represents a meaningful gain in reliability for accuracy switch cost and a meaningful loss in reliability for the latency switch cost in the presence of higher error rates, as compared with the previous experiment. Two of the proposed scores demonstrated enhanced reliability—the bin score with a reliability of $.831$ and the rate residual score with a reliability of $.723$. Both alternate scores met the reliability standards for research and developmental purposes, whereas the inverse efficiency score failed to do so with a reliability of $.630$.

Relationships with antisaccade analogue task performance

Table 5 provides the correlations between the task-switching scores and the antisaccade analogue scores. The correlation matrix shows that the relationships between the latency switch-cost scores and the proposed task-switching scoring

methods were moderate. Importantly, the relationships between the proposed task-switching scoring methods and the accuracy switch-cost scores were strong, indicating that the alternate scores successfully incorporated variation due to differences in accuracy switch costs.

All of the scoring methods had undetectable relationships with the antisaccade analogue scores, potentially due to a lack of power given the sample size.

Assessing a switch and congruency interaction

The Cohen's *d* effect sizes and mean scores for congruent and incongruent trials for each scoring method are provided in Table 7. Although each scoring method revealed an interaction between switch and congruency, the effect size of the interaction was larger for the rate residual, bin, and inverse efficiency switch-cost scores than for the latency and accuracy switch-cost scores. This supports the hypothesis that the alternate scoring methods had more power to detect the interaction effect because they combined two sources of information (latency and accuracy costs) into one score. This pattern was particularly important, given the higher error rates on the task-switching task.

Discussion

Across two experiments, we examined three alternative scoring methods to the traditional latency switch-cost score, which takes reaction time, but not accuracy, into account, and the traditional accuracy switch-cost score, which takes accuracy, but not reaction time, into account. All of the alternative scoring methods incorporate latency and accuracy information into a single score. Comparisons of these scores with the traditional methods, with respect to their reliability and internal and external validities, revealed several advantages of two of the proposed scoring methods (the rate residual and bin scoring methods). These advantages included superior reliability and strengthened relationships with an independent measure of executive control. Reliability differences were even more striking in Experiment 2—where a variable response deadline induced a higher error rate on the task-switching task—with reliabilities for the bin scores remaining in the .80 range, whereas traditional switch cost scores dropped to the upper .40s. In both experiments, the rate residual and bin scores were also better able to detect interactions between task switch and congruency, as indicated by larger effect sizes than for the latency, accuracy, and inverse efficiency switch-cost scores. Overall, the results of the present study support the use of either the rate residual or the bin scores as measures of task-switching ability that systematically account for switch costs to both accuracy and response latency.

Across both experiments, the three alternative scoring methods provided the most consistently reliable scores, whereas the least reliable was the accuracy switch-cost scores. Although the reliability of the latency switch-cost scores was satisfactory in Experiment 1, its reliability plummeted in Experiment 2 when participants demonstrated a much higher error rate. This indicates not only that the alternative scoring methods—particularly the rate residual and bin methods—better incorporate both latency and accuracy information, but also that these scoring methods are more robust to variations in the specific research paradigm employed within a given experiment. Indeed, the high reliabilities of our proposed scores across multiple paradigm variations give them an advantage over the scores with lower reliabilities regardless of the context of use (e.g., individual-difference research, experimental research on component processes of task switching). The error variance in the scores with lower reliabilities—such as the latency and accuracy switch costs—will reduce their validity as measures of task-switching ability, while also attenuating their relationships with other variables. This is especially true of the accuracy switch-cost scores with their unacceptably low level of reliability, a pattern that is only somewhat attenuated as error rates increase and, importantly, is accompanied by a decrease in the reliability of latency switch costs. This is likely due to the fact that higher error rates necessarily lead to fewer observations for latency switch cost scores, since errors are typically excluded from such scores. That is, separately analyzing latency and accuracy data can, in fact, harm the reliability and (potentially) validity of measurement of task performance. In contrast, the proposed scoring methods mitigate this risk by successfully incorporating both accuracy and latency into a unified score. These results suggest that researchers employing task designs where participants tend to make errors would benefit from employing rate residual or bin scores.

Experiment 1 demonstrated that the rate residual and bin scores were better equipped than the traditional methods to handle variation in both latency and accuracy switch costs. Variance in rate residual and bin scores for participants with similar latency switch-cost scores could be explained by differences in those participants' accuracy switch costs. Experiment 2 also demonstrated that the rate residual and bin scores captured accuracy switch costs with significant and moderate-to-strong correlations between the two alternate scores and accuracy switch-cost scores. This means that different levels of task-switching ability could appear as switch costs in reaction time, accuracy, or a combination of both and still be reflected by the rate residual and bin scores.

In addition to the advantages in reliability, the stronger relationship between the antisaccade analogue scores and task-switching performance provided a tangible demonstration of the most critical benefit of the rate residual and bin scoring methods: they provide superior measurement of the

underlying construct. As two measures of executive control abilities, performance on the task-switching and antisaccade analogue tasks was predicted to be moderately related. The strength of the relationship between traditional latency switch-cost scores and antisaccade analogue scores was similar to those found by Miyake et al. (2000), Friedman and Miyake (2004) and Friedman et al. (2006), who found correlations of .17, .25, and .23, respectively, between antisaccade analogue and similar task-switching tasks. To increase the likelihood of detecting a small effect such as this, task-switching scoring methods should incorporate all sources of variance that reflect task-switching ability. Although the relationships were weak, the rate residual score and bin scores were more highly related to the antisaccade analogue scores. In Experiment 1, correlations were larger in magnitude than those found with the latency switch-cost scores. However, we note that this benefit to validity was not consistently found in Experiment 2 involving higher error rates. Curiously, none of the correlations between the task-switching and antisaccade analogue scores reported in Experiment 2 were statistically significant, which is unexpected in light of published results (e.g., Miyake et al., 2000) and should be followed-up with experiments intended to compare these variables at the construct level (i.e., using latent variable analysis). Some correlations reported in Experiment 2 are relatively large, but the pattern appears random; our interpretation is that these are spurious due to small sample size.

Finally, the switch-by-congruency interaction effect analyses showed that these two alternate scoring methods were able to detect larger interaction effects than either the latency or accuracy switch costs alone. In particular, both the rate residual and bin scores consistently outperformed the traditional switch-cost scores, as well as the inverse efficiency switch-cost score, revealing larger effect sizes in both experiments. These results show that these two proposed scores are able to meaningfully combine information about latency and accuracy switch costs to provide a score that more powerfully reflects differences in task-switching ability. Researchers who utilize the rate residual or bin scores will be able to better reflect differences in performance costs across participants and experimental conditions. This could be especially important in studies with low power due to a small sample size and/or a focus on an effect that is small in magnitude and, therefore, more difficult to detect.

Taken together, the pattern of results across both experiments indicates that the use of rate residual or bin scoring may more accurately measure the underlying task-switching construct by capturing an additional, construct-relevant source of variance: accuracy switch-cost differences. Moreover, these two proposed scoring methods provided superior measurement qualities relative to the proposed inverse efficiency scoring method, although the inverse efficiency method did offer improvements over the traditional latency and accuracy

switch-cost scores. Given the importance of scoring metrics for both reliability and validity, future research should compare the two proposed methods with the inverse efficiency scoring method across a wider range of experimental paradigms. But it is clear that any of the three alternative methods is superior to the traditional scoring methods, and therefore researchers are encouraged to employ the rate residual or bin scoring methods in order to provide more precise measurement of task-switching performance.

It is critical to recall that these alternative scoring methods capture variance in performance from the very same task and do not require modifications to the research design. Indeed, the proposed scoring methods could be applied to existing data (sample scoring code provided as [supplemental material](#)) or be implemented alongside the traditional latency switch-cost score (as we have done here) in future studies to confirm that their more robust results generalize to other samples and research designs.

This study focused on two specific variants of the task-switching paradigm. Some task variants may measure task-switching ability better than others and/or tap into different aspects of task-switching ability in a way that is important to theory development. Indeed, a popular line of research is to compare different variants of the task to test specific predictions regarding the underlying processing mechanisms supporting performance. For example, previous research examining different cue–target intervals has found that longer cuing intervals typically reduce the magnitude of switch costs sharply (Meiran, 1996; Meiran, Chorev, & Sapir, 2000). The present results suggest that either of the proposed scoring methods could enhance measurement of task performance at different cuing intervals. Nonetheless, since different cuing intervals can lead to different patterns of results, we note that further research could be fruitful in determining whether varying the cue interval differentially affects the reliability and validity of the alternate scoring methods.

Previous research suggests that the specific task-switching variant can have significant consequences on the measurement of task-switching performance. For example, the alternating runs switching design used here may lead to inflated latency switch costs, due to the presence of a “restart cost” associated with the start of a new run (e.g., Altmann, 2007). Indeed, this is a serious issue to be considered by researchers studying task-switching performance. The present results indicate that researchers should consider both the task design and the scoring methods to optimize measurement of task-switching performance.

The goal of this article was to identify scoring methods that best measure task-switching performance to provide more reliable and valid indicators of task-switching ability across variants of the paradigm, rather than specific to a particular experimental design. The three scoring methods were selected to attempt to incorporate both accuracy and latency

information into one indicator of task performance. Specifically, we set the parameters of the scoring methods—particularly for the bin scores—to integrate the traditional cost for slower responses (i.e., longer latencies) with a cost for inaccurate responses (i.e., a penalty for errors). There may be other methods for penalizing inaccurate responses, and future research might compare possible alternatives with the methods proposed here. These scoring methods have utility both for researchers employing individual-difference methods—where a single indicator of task-switching ability is needed—and for those examining the effects of experimental manipulations on task performance. Future research could examine whether these scoring methods show similar benefits to task performance measurement across other variants to determine the robustness of these methods.

Conclusions

We examined alternatives to the traditional latency switch-cost scoring method across two variants of the task-switching paradigm. Results indicate that latency and accuracy switch-cost scores are less reliable than the alternative methods and, critically, are negatively affected by ignoring costs in accuracy and latency, respectively. Rate residual and bin scores provided better alternatives that were consistently reliable and more capable of capturing both accuracy and latency costs in a single score, thereby enhancing validity of construct measurement. Unlike the rate residual and bin scores, the inverse efficiency switch-cost score failed to consistently outperform latency and accuracy switch costs in our validity comparisons. Researchers should implement either rate residual or bin scores to improve measurement of task-switching ability, to optimize the likelihood of detecting relationships between task-switching performance and other variables of interest and to optimize the likelihood of detecting interaction effects within the task—all of which are critical to advances in theoretical models of task switching and executive control functions more broadly.

Acknowledgments This research was supported by the University of Maryland Center for Advanced Study of Language with funding from the United States Government. The authors would like to thank the members of the CASL project teams and DLIFLC for their efforts in data collection and Iring Koch, Kim-Phuong Vu, and an anonymous reviewer whose feedback on an earlier draft provided direction for strengthening this work. Portions of this research were presented at the 51st Annual Meeting of the Psychonomic Society, November 18–21, 2010.

References

- Altmann, E. M. (2007). Comparing switch costs: Alternating runs and explicit cuing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(3), 475–483.
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296–322.
- Bruyer, R., & Brysbaert, M. (2011). Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (IES) a better dependent variable than the mean reaction time (RT) and the percentage of errors (PE)? *Psychologica Belgica*, 51(1), 5–13.
- Bunting, M. F., Bowles, A. R., Campbell, S. G., Linck, J. L., Mislavy, M. A., Jackson, S. R., ... Doughty, C. J. (2010). *Reinventing DLAB: Potential new predictors of success at DLIFLC* (CASL Technical Report TTO 82114). College Park, MD: University of Maryland Center for Advanced Study of Language.
- Crocker, L., & Algina, J. (2006). *Introduction to classical & modern test theory*. Mason, OH: Thomson Wadsworth.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- De Jong, R. (2000). An intention-activation account of residual switch costs. In S. Monsell & J. Driver (Eds.), *Attention and performance XVIII* (pp. 357–376). Cambridge, MA: MIT Press.
- Einöther, S. J. L., Martens, V. E. G., Rycroft, J. A., & De Bruin, E. A. (2010). L-Theanine and caffeine improve task switching but not intersensory attention or subjective alertness. *Appetite*, 54, 406–409.
- Fagot, C. (1994). *Chronometric investigations of task switching*. (Doctoral dissertation, University of California, San Diego). Retrieved from http://www.pashler.com/Articles/techreport/Fagot_Dissertation1994.pdf
- Fidelman, C., & Kenyon, D. (2009). *Final report on measures proposed for a Hi-LAB GDS*. Washington, DC: Center for Applied Linguistics.
- Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent variable analysis. *Journal of Experimental Psychology: General*, 133, 101–135.
- Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological Science*, 17(2), 172–179.
- Gopher, D., Armony, L., & Greenspan, Y. (2000). Switching tasks and attention policies. *Journal of Experimental Psychology: General*, 129, 308–339.
- Gupta, R., Kar, B. R., & Srinivasan, N. (2009). Development of task switching and post-error-slowness in children. *Behavioral and Brain Functions*, 5, 38.
- Hallett, P. E. (1978). Primary and secondary saccades to goals defined by instructions. *Vision Research*, 18, 1279–1296.
- Jamadar, S., Michie, P., & Karayanidis, F. (2009). Sequence effects in cued task switching modulate response preparedness and repetition priming processes. *Psychophysiology*, 47(2), 365–386.
- Jamadar, S., Michie, P., & Karayanidis, F. (2010). Compensatory mechanisms underlie intact task-switching performance in schizophrenia. *Neuropsychologia*, 48, 1305–1323.
- Kiesel, A., Stenhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching – A review. *Psychological Bulletin*, 136, 849–874.
- Koch, I., & Allport, A. (2006). Cue-based preparation and stimulus-based priming of tasks in task switching. *Memory & Cognition*, 34, 433–444.
- Lien, M. C., Ruthruff, E., Remington, R. W., & Johnston, J. C. (2005). On the limits of advance preparation for a task switch: Do people prepare all the task some of the time or some of the task all of the time? *Journal of Experimental Psychology: Human Perception and Performance*, 31(2), 299–315.
- Linn, R. R., & Slinde, J. A. (1977). The determination of the significance of change between pre- and posttesting periods. *Review of Educational Research*, 47(1), 121–150.
- Logan, G. D. (2003). Executive control of thought and action: In search of the wild homunculus. *Current Directions in Psychological Science*, 12, 45–48.
- Meier, B., Woodward, T. S., Rey-Mermet, A., & Graf, P. (2009). The bivalency effect in task switching: General and enduring. *Canadian Journal of Experimental Psychology*, 63(3), 201–210.

- Meiran, N. (1996). The reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1423–1442.
- Meiran, N. (2000). Modeling cognitive control in task-switching. *Psychological Research*, *63*, 234–249.
- Meiran, N., Chorev, Z., & Sapir, A. (2000). Component processes in task switching. *Cognitive Psychology*, *41*, 211–253.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 48–100.
- Monsell, S. (2003). Task switching. *Trends in Cognitive Science*, *7*, 134–140.
- Monsell, S., Azuma, R., Eimer, M., Le Pelley, M., & Strafford, S. (1998). *Does a prepared task switch require an extra (control) process between stimulus onset and response selection?* Poster presented at the 18th International Symposium on Attention and Performance. Windsor, England: Great Park.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill, Inc.
- Pachella, R. G. (1974). The interpretation of reaction time in information processing research. In B. Kantowitz (Ed.), *Human information processing* (pp. 41–82). Potomac, MD: Lawrence Erlbaum.
- Poljac, E., Simon, S., Ringlever, L., Kalcik, D., Groen, W. B., Buitelaar, J. K., & Bekkering, H. (2010). Impaired task switching performance in children with dyslexia but not in children with autism. *The Quarterly Journal of Experimental Psychology*, *63*(2), 401–416.
- Prior, A., & Gollan, T. H. (2011). Good language-switchers are good task-switchers: Evidence from Spanish-English and Mandarin-English bilinguals. *Journal of the International Neuropsychological Society*, *17*, 1–10.
- Prior, A., & MacWhinney, B. (2010). A bilingual advantage in task switching. *Bilingualism: Language and Cognition*, *13*(2), 253–262.
- Rabbitt, P. (1997). Introduction: Methodologies and models in the study of executive function. In P. Rabbitt (Ed.), *Methodology of frontal and executive function* (pp. 1–38). Hove, England: Psychology Press.
- Rogers, R. D., & Monsell, S. (1995). The cost of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology*, *70*, 207–231.
- Rubin, O., & Meiran, N. (2005). On the origins of the task mixing cost in the cuing task-switching paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1477–1491.
- Rubinstein, J. S., Meyer, D. E., & Evans, E. (2001). Executive control of cognitive processes in task switching. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 763–797.
- Saeki, E., & Saito, S. (2009). Verbal representation in task order control: An examination with transition and task cues in random task switching. *Memory & Cognition*, *37*(7), 1040–1050.
- Salthouse, T. A., Fristoe, N., McGurthy, K. E., & Hambrick, D. Z. (1998). Relation of task switching to speed, age, and fluid intelligence. *Psychology and Aging*, *13*, 445–461.
- Samavatyan, H., & Leth-Steensen, C. (2009). The time course of task switching: A speed-accuracy trade-off analysis. *Memory & Cognition*, *37*(7), 1051–1058.
- Siemsen, E., Roth, A., & Oliveira, P. (2010). Common method bias in regression models with linear, quadratic, and interaction effects. *Organizational Research Methods*, *13*, 456–476.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295.
- Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan & F. Restle (Eds.), *Cognitive theory* (Vol. 3, pp. 200–239). Hillsdale, NJ: Erlbaum.
- Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. Cambridge: Cambridge University Press.
- Unsworth, N., Schrock, J. C., & Engle, R. W. (2004). Working memory capacity and the antisaccade task: Individual differences in voluntary saccade control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1302–1321.
- Vandierendonck, A., Liefvooghe, B., & Verbruggen, F. (2010). Task switching: Interplay of reconfiguration and interface control. *Psychological Bulletin*, *136*(4), 601–626.
- Was, C. A., & Woltz, D. J. (2007). Reexamining the relationship between working memory and comprehension: The role of available long-term memory. *Journal of Memory and Language*, *56*(1), 86–102.
- Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*(1), 67–85.
- Woltz, D. J., & Was, C. A. (2006). Availability of related long-term memory during and after attention focus in working memory. *Memory and Cognition*, *33*(3), 668–684.
- Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, *19*, 149–154.