aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2016, Vol. 11(1) 133–155

DOI: 10.1177/1745691615596990
pps.sagepub.com
SAGE

Combining Reaction Time and Accuracy: The Relationship Between Working Memory Capacity and Task Switching as a Case Example

Christopher Draheim, Kenny L. Hicks, and Randall W. Engle
Georgia Institute of Technology

Abstract

It is generally agreed upon that the mechanisms underlying task switching heavily depend on working memory, yet numerous studies have failed to show a strong relationship between working memory capacity (WMC) and task-switching ability. We argue that this relationship does indeed exist but that the dependent variable used to measure task switching is problematic. To support our claim, we reanalyzed data from two studies with a new scoring procedure that combines reaction time (RT) and accuracy into a single score. The reanalysis revealed a strong relationship between task switching and WMC that was not present when RT-based switch costs were used as the dependent variable. We discuss the theoretical implications of this finding along with the potential uses and limitations of the scoring procedure we used. More broadly, we emphasize the importance of using measures that incorporate speed and accuracy in other areas of research, particularly in comparisons of subjects differing in cognitive and developmental levels.

Keywords

task switching, working memory capacity, executive function, switch cost, measurement

Imagine yourself at home writing a lengthy e-mail when your phone rings; it is a friend telling you to turn on the news because of an interesting developing story. As you are talking to him, your dog begins barking at the door. You briefly disengage from the phone conversation to let your dog outside and change the television channel. Just then, your wife enters the room and asks who is on the phone. You respond, and she asks you to say hello for her. After obliging her, you return to chatting with your friend for a few minutes before sitting down to watch the news, while continuing to write the e-mail during the commercials.

A scenario such as the one described is by no means extraordinary or especially taxing, but consider how many occasions you had to disengage from your current activity to engage in another one that demanded immediate attention and then re-engage in the previous activity to pick up where you were when the interruption occurred. Your initial goal was simply to write an e-mail, but distractions arose, as they commonly do throughout the day. You properly attended to and responded to these distractions, but you also had to return to the mind-set of your initial goal. You had to remember that you were writing an e-mail and then the purpose of the email, what you were writing and thinking about in the e-mail, and some of your other thoughts before the interruption occurred. This ability to allocate attentive resources to several tasks sequentially and fluently reallocate attentive resources from one task to another is known as *task switching*, and it is an important higher-order ability.

In this example, an inability to attend to the distractions while switching back to your initial goal would not come at a high cost. However, as the individual tasks become more demanding, critical, and stressful, task switching becomes both more difficult and more important. Instead of imagining trying to write an e-mail on an ordinary morning, imagine being a soldier on the field of

Corresponding Author:
Christopher Draheim, School of Psychology, Georgia Institute of Technology, Atlanta, Georgia 30332
E-mail: cdraheim3@gatech.edu

battle having to listen to commands, read a map of the enemy position, and calculate the coordinates of an artillery strike, or imagine something most people have experienced—trying to carry on a phone conversation while driving in busy traffic. In these instances, being unable to appropriately switch between the two tasks could come at a very high price. Furthermore, whether you are a research professor, a short-order cook, a secretary, a stay-at-home parent, or a professional athlete, task switching is in some way relevant to your ability to perform your job. For these reasons, task switching is an important construct in psychology.

In the following sections, we introduce the theory and measurement of both task switching and working memory. We also tie working memory and task switching together, discussing the discrepancy between the theoretical relationship between the two constructs and empirical results from studies of this relationship. We provide an explanation and solution for this discrepancy, focusing on the nature of how task switching is measured, and what results indicate about the relationship between task switching and higher-order cognition. We then consider constructs other than task switching, because scoring issues (specifically, a lack of using both accuracy and RT) are problematic in other areas of psychology as well. Last, we explore the qualities of the procedure we used to score the task-switching data and urge researchers to exercise caution in choosing which dependent variable they use to represent their data.

Task Switching

History and measurement

The scientific study of task switching goes back to the 1920s (Jersild, 1927). Although task switching was not commonly studied during the information-processing era, a resurgence of interest in the mid-1990s brought about the introduction of new paradigms and techniques for assessing it (e.g., Allport, Styles, & Hsieh, 1994; Meiran, 1996; Rogers & Monsell, 1995). Advancements from research obtained with these new tools, along with the recognition of the ubiquity and importance of task switching, have led to task switching becoming a popular way to study cognitive control and executive functioning (e.g., Altmann & Gray, 2008; Logan, 2004; Miyake et al., 2000; Oberauer, Süß, Wilhelm, & Wittman, 2003).

In a typical task-switching experiment, subjects must alternate between making one of two (or more) simple judgments on sequentially presented trials. The judgments involve classifying a different attribute of the same set of stimuli. For example, the subjects might be shown a number along with a cue indicating they are to judge whether the number is even or odd or if the number is larger or smaller than 5. It is not surprising to find that subjects are generally slower and more prone to committing errors on trials in which they must make a different judgment from the previous trial (*switch trial*) as opposed to when they must make the same judgment as the previous trial (*nonswitch trial*). This disruption in response occurs even when there is ample warning and time to prepare for the upcoming switch (e.g., Allport et al., 1994; Meiran, 1996), and even with highly practiced subjects (Stoet & Snyder, 2007), indicating that this is a highly robust effect.

Task switching is often assessed in terms of *latency switch cost*—the amount of slowing that occurs on trials in which switching is required. Two types of latency switch costs can be calculated—local and global. For local switch costs, each subject's performances on switch and nonswitch trials in mixed blocks (i.e., blocks that contain both switch and nonswitch trials) are compared, and costs are calculated as the difference between the subject's mean RT on switch trials and his or her mean RT on nonswitch trials. Thus, a larger value indicates slower RT on switch trials relative to nonswitch trials.[1] For global switch costs, also called *mixing costs*, performance on nonswitch trials in mixed blocks is compared with that on nonswitch trials in pure blocks (i.e., blocks in which every trial is the same task, and thus no switching is required). Local switch costs are associated with processes involved in the actual execution of the task switch, whereas global switch costs are hypothesized to arise from the task ambiguity in mixed block conditions and reflect the retrieval and representation of the goals involved in task switching (e.g., Chevalier et al., 2012; Rubin & Meiran, 2005). In terms of the example described at the beginning of this article, local switch costs would be reflected in any slowing that occurred when you began talking to your friend on the phone again and had to retrieve what you were talking about before letting the dog outside, turning on the television, and acknowledging your wife. Global switch costs would be reflected in any additional slowing that occurred when you were writing out the e-mail during the commercials versus when you were initially writing out the e-mail undisturbed by external distractions.

Paradigms

We will discuss only a few of the numerous methods used to induce the switch cost here. Jersild's (1927) approach was to obtain the baseline amount of time it took the subject to complete a list containing items that did not require switching (e.g., adding 6 to each number) and subtract that from the amount of time it took to

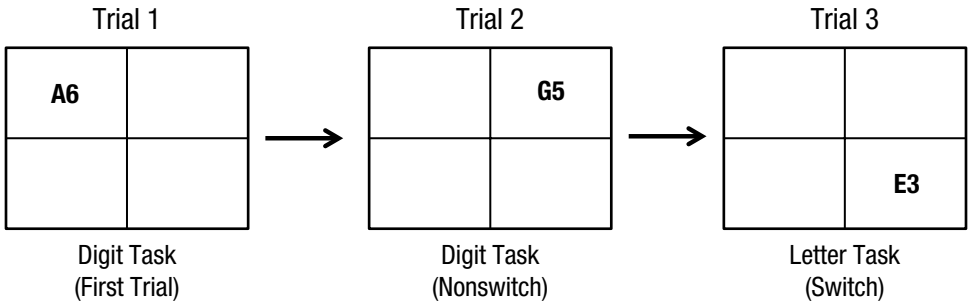

Fig. 1. Example of the Rogers and Monsell (1995) alternating-runs paradigm. The subject is asked to judge whether the number is even or odd if the stimulus appears in the top half of the grid or whether the letter is a vowel or consonant if the stimulus appears in the bottom half. The stimulus predictably moves about the grid in a clockwise fashion; thus, every even trial is nonswitch, and every odd trial is switch.

complete a list in which switching was required for every item (e.g., add 6 to a number, then subtract 3, then add 6, and so forth). This approach, while elegantly simple, is problematic in that local and global switch costs are confounded. The *alternating-runs* procedure was introduced by Rogers and Monsell (1995) as a way to disentangle the local switch costs from the global. In this design, both switch and nonswitch trials are mixed within the same block, and the trials follow a predictable sequence in which every other trial is a switch trial (see Fig. 1). Another approach is the *cueing* procedure, in which the trials occur in a pseudo-random order, and the subject is externally cued as to which task to perform on a trial-by-trial basis (e.g., a heart or a cross appearing just prior to stimulus onset, Meiran, 1996; see Fig. 2). An advantage of the cueing procedure is that it allows more experimental control because the cue-stimulus and response-cue intervals can be independently manipulated. These procedures have been criticized in terms of their ecological validity because subjects are not free to switch of their own volition, and thus, the *voluntary* procedure has also become a popular method to assess endogenous switching processes (Arrington & Logan, 2004; Mayr & Bell, 2006).

Theories

Regardless of the paradigm employed, a memory representation of the appropriate configuration of rules for performing each task (a *task set*) has to be maintained in easily accessible form and retrieved to properly switch tasks. Various processes involving task sets have been proposed as the source of switch costs, and there are two popular theoretical accounts. The first, reconfiguration theory, posits that switch costs arise from a task-set reconfiguration in which the cognitive system has to reconfigure itself to properly execute the switch (e.g., returning to the mental state of the phone conversation with your friend after you have let the dog out and turned on the television; Rogers & Monsell, 1995). The other, interference theory (also referred to as *task-set inertia*), posits that switch costs reflect proactive interference from the previously active but now irrelevant task set (e.g., thinking about the news story while trying to continue writing the e-mail; Allport et al., 1994). Although researchers usually support one of these perspectives over another, a strong case can be made that they have much in common and are not mutually exclusive (e.g., Vandierendonck, Liefooghe, & Verbruggen, 2010). Because settling the

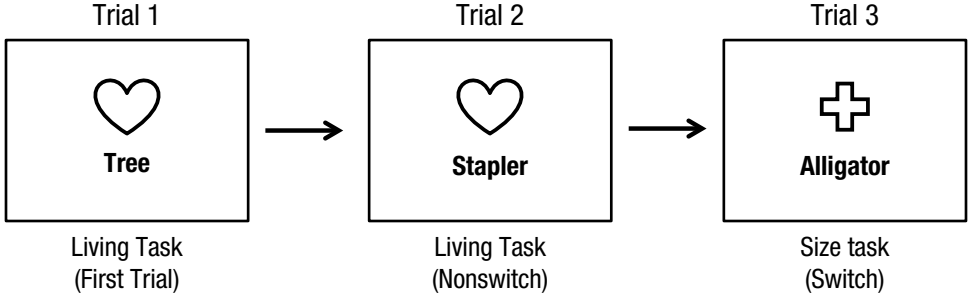

Fig. 2. Example of the task-cueing paradigm. The subject judges whether the object represented is living or nonliving if the heart cue appears or whether the object represented is larger or smaller than a referent (e.g., soccer ball) if the cross cue appears. The cues appear in pseudo-random fashion; hence, the subject does not know when switch trials are coming.

debate between these two theories goes beyond the scope of this article, we discuss task switching from the perspective of reconfiguration theory while also keeping in mind that it and interference theory are likely compatible with one another (for review, see Kiesel et al., 2010; Monsell, 2003; Vandierendonck et al., 2010).

Working Memory

Introduction to working memory

Working memory is the domain-general ability to simultaneously maintain, process, and manipulate chunks of goal-relevant information and is an important aspect of cognitive control. As opposed to long-term memory, working memory has limitations in terms of how much information can be maintained at any given time. Therefore, it is typically measured in terms of capacity—estimated to be between three and five chunks of information in normal functioning individuals (Cowan, 2001). Working memory capacity (WMC) is a critical construct in many areas of psychology because it has been shown to predict a wide range of cognitively complex, real-world behaviors. Specifically, individuals with larger WMC have been shown to be better at following directions (Engle, Carullo, & Collins, 1991), multitasking (Hambrick, Oswald, Darowski, Rench, & Brou, 2010), language learning (e.g., Baddeley, Gathercole, & Papagno, 1998), language comprehension (e.g., Daneman & Merikle, 1996), attentional control (e.g., Kane, Bleckley, Conway, & Engle, 2001), and reasoning (e.g., Kyllonen & Christal, 1990). More importantly, WMC has been shown to share a substantial amount of variance with fluid intelligence (Gf), which is the ability to reason in novel situations (Ackerman, Beier, & Boyle, 2005; Engle, Tuholski, Laughlin, & Conway, 1999; Kane, Hambrick, & Conway, 2005; Oberauer, Schulze, Wilhelm, & Süß, 2005).

Measuring working memory

Working memory can be studied through either experimental (group) or differential (individual differences) perspectives. With the experimental approach, researchers often impose a form of working memory load (e.g., a dual task or a form of interference) in one condition and compare performance in the load condition with that in a nonload control condition (e.g., Baddeley et al., 1998). If performance is lower in the high load condition, then the conclusion is drawn that working memory is important in being able to perform the task. With the differential approach, the assumption is that any relationship between WMC tasks and other cognitive tasks may mean that working memory is important in that other cognitive task. The differential approach relies on correlations, which permits the investigation into the underlying structure of cognition using latent variable analyses.

For example, a common way to measure WMC is with complex span tasks, the first of which were the reading span (Daneman & Carpenter, 1980), counting span (Case, Kurland, & Goldberg, 1982), and the operation span (Turner & Engle, 1989). Since then, spatial complex span tasks such as the rotation span (Shah & Miyake, 1996) and symmetry span (Kane et al., 2004) have also been introduced. Although the stimuli of these tasks vary, they share a similar design: The subject performs a simple processing task (e.g., basic arithmetic or a symmetry judgment) and then is presented with a stimulus to maintain in short-term memory (e.g., a letter or particular cell within a grid). After a certain number of presentations (typically ranging from two to seven), the subject is asked to recall the stimuli in the order in which they were presented. This process is repeated, and at the end of the task, a span score is calculated that reflects how many stimuli the subject recalled in the correct order. These tasks have been shown to have both high reliability and validity (see Conway et al., 2005, and Redick et al., 2012, for reviews). However, like most cognitive contructs, WMC is best measured with multiple tasks because of the presence of construct-irrelevant and task-specific variance in any one particular task. Figure 3 illustrates two complex span tasks—the operation span and symmetry span.

Arguably, the most direct way to study the relationship between working memory and other cognitive constructs is to measure WMC through a variety of tasks and use the latent variable approach to relate WMC to other constructs. However, such methods require both a large sample size and a large battery of tasks. As previously mentioned, studies using these methods have routinely shown a strong causal relationship between WMC and other higher order constructs and real-world abilities. Furthermore, understanding the relationships can illuminate cognitive impairments and learning difficulties because executive attention, a critical component of WMC (see Engle, 2002), is also important for problem solving, reasoning, and learning (i.e., abilities afforded by Gf). Additionally, cognitive deficits and clinical pathologies (e.g., attention-deficit/hyperactivity disorder, schizophrenia, and Alzheimer's disease) are all strongly associated with deficits in WMC.

The Relation Between Task Switching and Working Memory

In task switching, it is presumed that multiple task sets cannot be simultaneously active, and thus successful switching requires both the deactivation of the old (now-irrelevant) task set and also the re-activation of the new

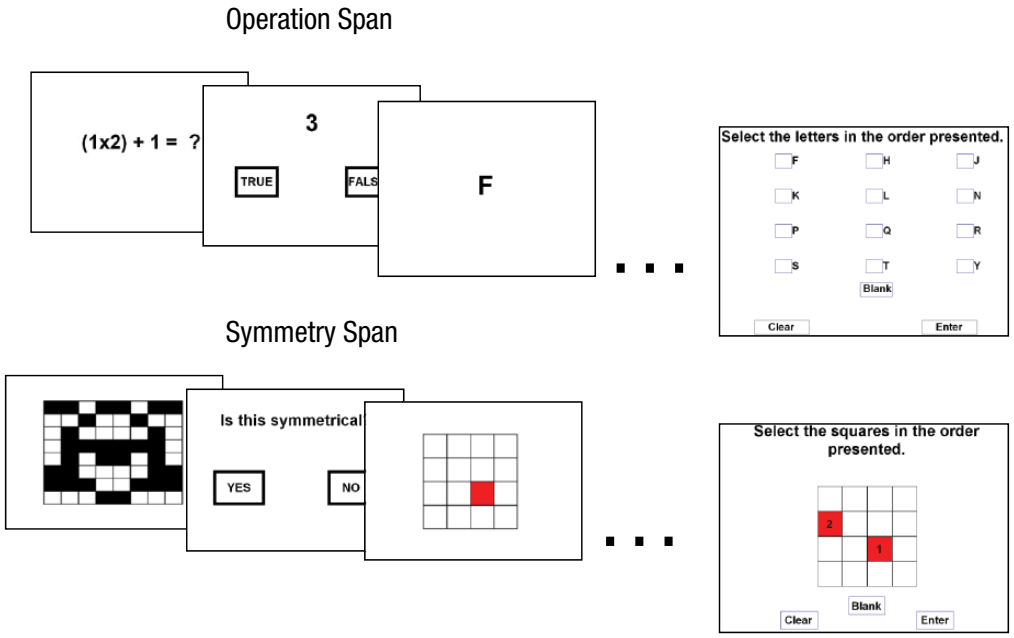

Fig. 3. Example of the operation-span and symmetry-span tasks. This figure shows one presentation of the processing and storage tasks and then the final recall screen seen by the subject after between two and seven such presentations. After recall, the process is repeated. For the operation span, three trials with set sizes of three, four, five, six, and seven are typically administered. For the symmetry span, three trials with set sizes of two, three, four, five, and six are typically administered. Adapted from Fig. 1 of "Working memory training may increase working memory capacity but not fluid intelligence," by T. L. Harrison, Z. Shipstead, K. L. Hicks, D. Z. Hambrick, T. S. Redick, & R. W. Engle, 2013, *Psychological Science, 24*, p. 2411. Copyright 2013 by Association for Psychological Science.

(now-relevant) set into working memory (Mayr & Keele, 2000; Monsell, 2003). The precise mechanisms of the reconfiguration process are debated, but the proposed functions include shifting attention between stimulus attributes, retrieving goal states and condition-action rules into working memory, and inhibiting elements of the prior task set while activating the new, appropriate task set (Monsell, 2003). These processes are related to working memory, and indeed some reconfiguration theorists have argued that task switching is completely mediated through working memory (Mayr & Kliegl, 2000; Rubinstein, Meyer, & Evans, 2001). Although researchers have implicated processes outside working memory (e.g., retrieval from long-term memory; Allport et al., 1994; Logan & Gordon, 2001) as also important, they mostly agree that working memory is critical in the ability to switch tasks. Specifically, task switching is considered to be a hallmark of executive control (e.g., Logan, 2004), and one would expect that task-switching performance and WMC would be highly correlated (see Kane, Conway, Hambrick, & Engle, 2007). Some evidence of this relationship comes from a series of studies (performed with experimental approaches) that have shown that the central executive is involved in task switching (Baddeley, Chincotta, & Adlam, 2001), that the phonological loop[2] is involved in the retrieval of task sets (Emerson & Miyake, 2003; Liefooghe, Vandierendonck, Muyllaert, Verbruggen, & Vanneste, 2005), and that task switching exacts a cost on working memory functioning (Liefooghe, Barrouillet, Vandierendonck, & Camos, 2008).

Our correlational study

Recently, we conducted a large-scale correlational study (n = 552) investigating the relationship among executive functions, WMC, and fluid intelligence (Shipstead et al., 2015). Task switching was included in this study, and we assessed it from two cueing tasks (category switch and letter-number switch) with a 5,000-ms response deadline. We assessed WMC using automated versions of the operation span, symmetry span, rotation span, and letter-number-sequence. We assessed Gf using Raven's advanced progressive matrices, letter sets, and number series. In regards to task switching, our hypothesis was that individuals with higher levels of WMC and Gf would perform better at task switching.[3]

To our surprise, the initial analysis of the data revealed a pattern of results that did not support this hypothesis. In fact, individuals with lower levels of WMC and Gf were better at task switching than those with higher levels. That is, individuals who scored lower on the WMC and Gf tasks showed smaller switching costs than

individuals who scored higher on WMC and Gf tasks. Table 1 shows the descriptive statistics for these data, and Table 2 displays the correlations among the WMC, Gf, and the task-switching tasks. The correlations (*r*s) between number switching and the four indicators of WMC ranged from −.22 to −.26, all statistically significant (α= .05).[4] The correlations (*r*s) between category switching and the four indicators of WMC ranged from −.08 to −.17, with all but one being statistically significant. WMC composite performance and task-switching composite performance, formed from the *z*-score averages of the tasks, correlated (*r*) at −.26 ($p < .001$), indicating that task switching and WMC shared almost 7% common variance, opposite from the direction predicted by theory. The same pattern was found between task switching and Gf, in which all zero-order correlations (*r*s) between tasks were significant (ranging from −.11 to −.33), and the composite Gf score correlated (*r*) −.32 with the composite task switching score ($p < .001$), sharing slightly over 10% variance in the opposite direction as predicted. In short, these results lead us to believe that having high WMC or Gf is actually detrimental to task-switching performance, a finding that is highly unintuitive and requires further investigation.

Other studies

Given the literature, perhaps we should not have been surprised at our results. It is a common finding that WMC is not related to task switching at either the individual task or latent level (e.g., Kiesel, Wendt, & Peters, 2007; Logan, 2004; Miyake et al., 2000; Oberauer, Süß, Wilhelm, & Wittman, 2003). The same can be said of the relationship between Gf and task switching, with the caveat that many of these studies are neuropsychological or developmental in nature and may not use the most psychometrically sound tasks (see Friedman et al., 2006).

For our purposes, the studies conducted by Miyake et al. (2000) and Oberauer et al. (2003) are particularly important. In these large-scale correlational studies, a latent variable approach was used to model the relationship of executive functions in a similar manner to that used in Shipstead et al. (2015), and both also included several measures of task switching. The conclusions made from these two studies were that the relation of task switching to other executive functioning tasks ranged from weak to moderate but that the relation of task switching to WMC was either very weakly or nonexistent.[5] Specifically, Miyake et al. reported correlations (*r*s) ranging from −.04 to .09 between the operation span and three different task-switching tasks (none of these correlations was statistically significant at α= .05). Oberauer et al. used both log-transformed local and global switch costs and reported correlations from −.07 to .23 between task switching and their six markers of working memory (with only three of these twelve correlations being significant at α = .05).

If we focus on the individual differences studies and take the findings from Miyake et al. (2000), Oberauer et al. (2003), and Shipstead et al. (2015) at face value, the

Table 1. Descriptive Statistics From the Initial Analysis of Shipstead, Harrison, and Engle (2014)

Task	*M*	*SD*	Range	Skew	Kurtosis	I.C.
WMC						
1. OSpan	54.18	15.51	.00 – 75.00	−.88	.20	.86[a]
2. SymSpan	26.65	9.05	3.00 – 42.00	−.43	−.49	.84[a]
3. RotSpan	24.63	9.77	.00 – 42.00	−.44	−.57	.87[a]
4. LNS	10.84	4.00	.00 – 23.00	−.01	.11	.85[a]
Gf						
5. Raven	8.69	3.91	.00 – 18.00	−.07	−.88	.82[a]
6. LetterSet	15.38	5.29	1.00 – 29.00	−.03	−.62	.84[a]
7. NumSeries	8.56	3.58	.00 – 15.00	−.22	−.86	.83[a]
Task switching						
8. NumSwitch	300.18	238.89	−519.65 – 1145.62	.23	.46	.73[b]
9. CatSwitch	233.65	179.00	−249.55 – 866.30	.71	.63	.63[b]

Note. $N = 552$. WMC = working memory capacity; OSpan = operation span; SymSpan = symmetry span; RotSpan = rotation span; LNS = letter number sequencing; Gf = fluid intelligence; Raven = Raven's advanced progressive matrices; LetterSet = letter sets; NumSeries = number series; NumSwitch = letter-number switch; CatSwitch = category switch.

[a]Internal consistency (I.C.) was calculated using Cronbach's α. [b]I.C. was calculated using a split-half procedure and was stepped up according to the Spearman-Brown prophecy formula.

Table 2. Zero-Order Correlations of the Working Memory Capacity, Fluid Intelligence, and Task-Switching Measures From the Initial Analysis of Shipstead, Harrison, and Engle (2014)

	WMC				Gf			Switching	
Task	1	2	3	4	5	6	7	8	9
1. OSpan	1.00								
2. RotSpan	.52*	1.00							
3. SymSpan	.53*	.68*	1.00						
4. LNS	.46*	.48*	.45*	1.00					
5. Raven	.50*	.59*	.52*	.51*	1.00				
6. LetterSet	.45*	.55*	.49*	.54*	.61*	1.00			
7. NumSeries	.54*	.55*	.55*	.52*	.65*	.68*	1.00		
8. NumSwitch	−.22*	−.24*	−.23*	−.26*	−.33*	−.30*	−.27*	1.00	
9. CatSwitch	−.17*	−.12*	−.08	−.12*	−.20*	−.11*	−.13*	.29*	1.00

Note. $N = 552$. WMC = working memory capacity; Gf = fluid intelligence; Ospan = operation span; RotSpan = rotation span; SymSpan = symmetry span; LNS = letter number sequencing; Raven = Raven's advanced progressive matrices; LetterSet = letter sets; NumSeries = number series; NumSwitch = letter-number switch; CatSwitch = category switch.
The dependent variable in letter-number switch and category switch was a latency switch cost. Correlations involving task switching and another measure were multiplied by −1 for ease of interpretation such that a positive correlation between any two variables suggests individuals who performed better on one task tended to also perform better on the other.
*$p < .05$.

only reasonable conclusion is that having a larger WMC does not facilitate, and possibly may even hinder, task-switching performance. This finding seems robust considering the diversity in methodologies employed in the studies across the three different labs. Miyake et al. (2000) used three different types of switching tasks (Jersild's list procedure, alternating runs, and cueing), and performance on these did not correlate with the operation span in a sample of 137 undergraduates in America. Oberauer et al. (2003) had a sample of 135 undergraduates in Germany perform four alternating-runs tasks (calculating both local and global switch costs) and found very weak correlations between performance on these tasks and their own working memory tasks (for more detail about these tasks, see Oberauer, Suß, Schulze, Wilhelm, & Wittmann, 2000). Shipstead et al. (2015) had a diverse sample of 552 individuals from two American universities and the community in Atlanta, Georgia, perform two cueing tasks with a 5,000-ms response deadline and found a negative correlation between task switching and performance on four WMC tasks (three being complex span tasks).

Interim Summary

Thus far, we have introduced the constructs of task switching and WMC. Theoretical accounts of these two constructs predict a fairly large and positive relationship between the two. However, past research presents a conundrum to researchers. On one hand, there is some evidence from experimental approaches that task switching is indeed negatively affected when a high working memory load is present in the task, suggesting that working memory at least partially mediates task switching. On the other hand, other experimental studies have failed to show this relation. Adding to the confusion, three large-scale studies in which the differential approach was used also yielded null or even negative results in terms of WMC and task switching being positively related. These findings have led researchers to call into question the theoretical accounts of task switching, and claim either that working memory is not involved in task-set reconfiguration or that task-set reconfiguration is not the source of switch costs (see Liefooghe, Barrouillet, Vandierendonck, & Camos, 2008). Additionally, the executive attention theory of WMC, a widely supported theory of working memory, has also been criticized on the basis of these findings (Kane et al., 2007; see Oberaurer, Süß, Wilhelm, & Sander, 2007).

In short, given that task switching is viewed as a prototypical executive function that requires processes common to working memory, the failure to consistently demonstrate a substantial relation between task switching and WMC is problematic for both working memory and task-switching theories and presents an intriguing challenge to researchers. In the following sections, we argue that these results are a product of how task switching is measured, introduce a solution to this

measurement issue, reanalyze data from Oberauer et al. (2003) and Shipstead et al. (2015) with an alternative scoring procedure to support our claim, and discuss both the broad and narrow implications of our results.

Why Latency Switch Costs Are Problematic

Despite the methodological differences in Miyake et al. (2000), Oberauer et al. (2003), and Shipstead et al. (2015), these studies do share one thing: The dependent variable was an RT-based cost score. Latency switch costs have been used to assess task switching since the introduction of Jersild's (1927) multilist procedure, in which the total RT to complete a pure list of problems (i.e., no switching involved) is subtracted from the total RT to complete a list of problems in which every trial requires switching. Latency switch costs continue to be the most frequently used measure of task-switching performance (whether using local or global costs), often without much consideration given to their measurement properties. As it turns out, there are at least two major reasons why researchers should be cautious in using latency switch costs. The first concern is methodological, and the second is psychometric.

The most readily apparent issue with latency switch costs is that accuracy is not taken into account, and thus major differences in accuracy rates can go undetected. Using a score in which accuracy is completely ignored is problematic both in group comparisons if any of the groups differ in accuracy and in differential approaches if there are individual differences in how subjects adjust their speed and accuracy against one another. Unless researchers specifically manipulate speed-accuracy tradeoffs, instructions for most cognitive tasks direct the subject to answer as quickly as possible while maintaining high accuracy. This instruction is designed to equate subjects in terms of how much emphasis they give to both speed and accuracy, but it is likely that these instructions alone are not sufficient and that individual differences will still emerge.[6] For example, some subjects may simply have a tendency to maintain high accuracy levels and will do so with the consequence of being slower, resulting in high latency switch costs. On the other hand, more hasty or impulsive subjects may have a tendency to produce more errors on switch trials while exhibiting low latency switch costs. Additionally, some subjects simply may be unable to make the appropriate speed-accuracy adjustment that is required for the particular task being performed, whereas others quickly and appropriately adjust to meet task demands.

Our argument here is that high-ability individuals do just that. They are more likely to adjust their speed to maintain accuracy than are low-ability individuals. Assessing task-switching performance with latency switch costs results in researchers being unable to differentiate between these different types of subjects and can make low-ability subjects who sacrificed accuracy to be quicker look as though they performed better than high-ability subjects who followed task instructions and maintained a high level of accuracy. This explanation is consistent with the findings of Unsworth and Engle (2008) in which WMC and Gf were not related to the amount of time it took to switch during continuous counting tasks but were related to the accuracy rates of these tasks, as low-ability individuals were much more error prone.

The other major issue with switch costs is that difference scores in general have such low reliability that some researchers have advised against using them in any circumstance (Cronbach & Furby, 1970; Edwards, 2001; Lord, 1963; Peter, Churchill, & Brown, 1993). The simplest formula for estimating reliability of a difference score is given by Guilford (1954) and Lord (1963):

$$\rho_{dd'} = \frac{\rho_{xx'} - \rho_{xy}}{1 - \rho_{xy}}.$$

In this formula, $\rho_{dd'}$ is estimated reliability of the difference score, $\rho_{xx'}$ is estimated reliability of the two component scores, and ρ_{xy} = correlation between the two component scores.

Thus, as the correlation between the two component scores increases, the reliability of the difference score decreases. Because difference scores are calculated as a within-subject variable and the two component scores are designed to measure a very similar process, the correlation between the component scores is expected to be high and positive (e.g., in Shipstead et al., 2015, nonswitch RT and switch RT were correlated (r) .89 in category switching and .90 in number switching). The result is a score with low reliability, ultimately restricting potential validity.[7] Additionally, difference score reliability can vary widely from task to task even within the same experiment. For example, in Miyake et al. (2000), the switch costs of one task had a reported internal consistency of .59, whereas another had a reported internal consistency of .91 (both assessed via Cronbach's α). In the initial analysis of Shipstead et al., the task-switching tasks had an internal consistency of .63 and .73 (assessed via split-half and stepped up with the Spearman-Brown prophecy formula), markedly lower than the internal consistency of the other tasks in this study. Although these estimates are low, they are likely upwardly biased such that the true reliability is lower than the internal consistency.[8] For example, task-switching tasks in Shipstead et al. correlated (r) at only .29 despite being

highly similar tasks, indicating that validity was likely attenuated.[9]

As a result of these issues, use of latency switch costs in task-switching studies can result in faulty conclusions and misguided theory. In a recent article, Hughes, Linck, Bowles, Koeth, and Bunting (2014) argued that using latency switch costs places heavy restrictions and limitations on theoretical developments from task-switching studies. They also argued that assessing latency switch costs and accuracy switch costs separately still fails to reveal relations and interactions because each score reflects only one cost at a time and both measures are difference scores that have low reliability. Like other researchers who have urged against using differences scores, they advised against the practice of using switch costs in task-switching paradigms. According to this analysis, the discrepant results found in studies of the relation between working memory and task switching might very well be due to issues with the dependent variable used to measure task switching rather than the theoretical underpinnings of either construct.

An Alternative to Latency Switch Costs: The Binning Procedure

Hughes et al. (2014) proposed three alternative scoring procedures designed to overcome the issues associated with latency switch costs. Among these alternatives is a rank-ordering binning procedure that combines speed and accuracy to form a single, comprehensive score of task-switching performance. Scores from this procedure are calculated in the following manner:

Step 1. Mean RTs on accurate nonswitch trials are calculated for each subject.

Step 2. The mean RT from Step 1 is subtracted from the RT for each subject's individual accurate switch trial. This procedure results in every accurate switch trial having a score that represents how fast the subject responded on that particular trial relative to his or her own average nonswitch RT. This is a within-subject comparison done for each subject.

Step 3. The scores from Step 2 for all subjects combined are rank ordered into deciles and assigned a bin value ranging from 1 to 10. The fastest 10% of scores (again, for all subjects as a group) is assigned a value of 1, the next 10% is given a value of 2, and so forth, until the slowest 10% of scores (i.e. the slowest responses for subjects relative to their own nonswitch RT) is given a value of 10. This procedure results in every accurate switch trial having a corresponding bin value ranging from 1 to 10. A trial with a value of 1 means that on that particular switch trial, the subject's response was quicker than 90% of all other responses for all subjects (comparing accurate switch trials to a particular subject's mean RT on all nonswitch trials).

Step 4. Inaccurate switch trials (which had been ignored up to this point) are assigned a bin value of 20, regardless of RT. Hence, inaccurate switch trials are given a value twice as high as the slowest accurate switch trial. At this point, data for each subject consist of each switch trial having a corresponding bin value, ranging from 1 to 10 or 20.

Step 5. A single bin score is computed for each subject by summing all of their respective bin values.

A smaller bin score for a subject indicates a combination of two things. First, on accurate switch trials, that subject's RT tended to be only slightly larger than his or her mean RT for nonswitch trials (i.e. low—potentially even zero or negative—latency switch costs) compared with those of other subjects. Second, the subject made fewer errors on switch trials than other subjects. Thus, this method incorporates both RT and accuracy data from the task into one comprehensive score and provides more information than traditional techniques in which one of the two measures is ignored or speed and accuracy are analyzed. This method was also shown to have high reliability in Hughes et al. (2014), whereas both latency and accuracy switch costs had low reliability.

Reanalysis of Study Data

Shipstead et al. (2015)

In an effort to test the robustness of this alternative method and answer questions from our own task-switching experiment, we reanalyzed data from Shipstead et al. (2015) using the binning procedure. As the reader will recall, this study had 552 subjects perform 50 cognitive tasks, including four tasks designed to measure WMC, three tasks designed to measure Gf, and two tasks designed to measure task-switching ability. The initial analysis of these data (using latency switch costs to measure task switching) revealed a surprising trend in which individuals who scored higher on the WMC and Gf tasks performed worse on the task-switching tasks (refer to Table 2).

When we applied the binning procedure to these data, the results were quite different. In regards to WMC, the correlation between the composite score on the WMC tasks and the composite score of the task-switching tasks changed dramatically both in terms of magnitude and direction. The initial analysis (with latency switch costs) revealed a correlation (r) of −.26 ($p < .001$)

between composite task-switching and WMC scores, whereas the reanalysis with the binning procedure revealed a correlation (*r*) of .49 ($p < .001$). Thus, scores from the binning procedure explained 24% of the variance in WMC, which is triple the amount explained when data were analyzed with latency switch costs, and the bin scores also were in the hypothesized direction. Thus, the binning procedure indicates quite clearly that individuals who performed better on the WMC tasks also performed better on the task-switching tasks when speed and accuracy are combined into a single metric. This conclusion could not have been made on the basis of latency switch costs or, for that matter, accuracy switch costs alone. Furthermore, this conclusion could not have been made on the basis of an analysis in which latency and accuracy switch costs were calculated separately; in such a analysis, the correlation between task switching measured by accuracy switch costs and performance on the WMC tasks effectively was zero, and the correlation between task switching measured by latency switch costs and performance on the WMC tasks was negative. Table 3 illustrates the comparison of different scoring techniques of the task-switching tasks in terms of zero-order correlations with the composite WMC score. Table 4 displays the descriptive statistics and reliability of data analyzed with either latency switch costs or the binning procedure. Not only were these correlations substantially different, but also the estimated reliability of the scores from the binning procedure was an improvement over latency switch costs (with an estimated .72 and .83 internal consistency, improvements from .63 and .73, respectively). The improvement in reliability also was manifest in the increase in the correlation (*r*) among the two task-switching tasks: .52 with the binning procedure versus .29 with use of latency switch costs. It is evident that using the binning procedure to analyze these data tells a very different story than when using switch costs based on latency, as the hypothesized positive correlation between task-switching performance and WMC emerges.[10] We will discuss these findings in more detail in some of the following sections.

Table 3. The Zero-Order Correlations of the Composite WMC Score and the Composite Task Switching Score Measured in Three Different Ways From Shipstead, Harrison, and Engle (2014)

Measure	1	2	3	4
1. WMC	1.00			
2. Latency switch cost	−.26*	1.00		
3. Accuracy switch cost	.01	−.08	1.00	
4. Bin score	.49*	.18*	.26*	1.00

Note. $N = 552$. WMC = composite score on the four working memory capacity tasks; Bin score = composite task switching score of the two tasks analyzed via the binning procedure. Correlations involving WMC were multiplied by −1 such that a positive correlation indicates that individuals who performed better on the WMC tasks also tended to perform better on the switching tasks.
*$p < .05$.

Oberauer et al. (2003)

For replication of the binning procedure across a new sample and different types of tasks, we also reanalyzed data from Oberauer et al. (2003). In this study, Oberauer et al. tested 135 University of Mannheim students on 24 tasks of executive functioning along with six working memory tasks. One of the executive functions, supervision, consisted of four Monsell-like alternating-runs tasks, each with 96 trials (50% being switch trials). The two other executive functions were (a) storage and processing and (b) coordination (see Oberauer et al., 2003, for a full description).

Table 5 shows the zero-order level correlations between the working memory tasks and composite task-switching performance scored by either latency switch costs or the binning procedure. Just as in Shipstead et al. (2015), binning these data led to a substantial increase in the relationship between task switching and WMC. All six of the working memory tasks correlated significantly with the task-switching composite score when data were analyzed via binning, whereas only two of the six were significant when data were analyzed with latency switch costs. In the test for differences between dependent correlations (Steiger, 1980), use of the binning procedure resulted in four out of the six correlations being significantly different from correlations found via use of latency switch costs. The greatest improvement came from forming composite working memory scores based on the six tasks; for example, a correlation (*r*) .37 between working memory and task switching was observed with the

Table 4. Descriptive Statistics of the Two Task-Switching Tasks in Shipstead, Harrison, and Engle (2014) Analyzed via Latency Switch Costs or the Binning Procedure

Task	*M*	*SD*	Skew	Kurtosis	I.C.
	Latency switch cost				
Category Switch	233.65	179.00	.71	.63	.63
Number Switch	300.18	238.89	.23	.46	.73
	Bin score				
Category Switch	374.45	103.25	.92	.54	.72
Number Switch	392.19	109.86	.87	.84	.87

Note. I.C. = Internal consistency. Internal consistency was estimated by Guttman split-half and stepped up according to the Spearman-Brown prophecy formula.

binning procedure, whereas a correlation of .20 was observed with the analysis of latency switch costs (these correlations were significantly different from one another, $p < .001$). Thus, performance on the working memory tasks shared almost 14% variance with task switching when analyzed via the binning procedure but only 4% when analyzed via analysis of latency switch costs (and the latter was not in the hypothesized direction). Table 6 shows a stepwise regression analysis predicting performance on the working memory tasks. Adding composite bin scores of the task-switching tasks as a predictor of WMC resulted in a significant improvement over the model in which latency switch costs were used; however, the opposite was not true. Therefore, task-switching bin scores explained unique variance in working memory above and beyond latency switch costs, but latency switch costs did not explain statistically significant unique variance over bin scores.

At the latent level, binning the task-switching data significantly improved the amount of variance shared by supervision and the other executive functions. Specifically, the amount of shared variance between the supervision and storage and processing factors above and beyond coordination tripled, as does the amount of shared variance between supervision and coordination above and beyond storage and processing. Figure 4 shows the original model from the study, and Figure 5 shows the same model when scores from the binning procedure were used as the dependent variable. These results demonstrate that the binning procedure captured a component of task-switching performance that is common with the other executive functions in the study, indicating that it measures an important aspect of executive functioning. The results of Shipstead et al. (2015) and Oberauer et al. (2003) provide support for the binning procedure being a robust scoring technique, as improvements over latency switch costs were observed in two data sets that had subjects with different demographics and tasks representing different types of task switching and WMC.

Incorporating RT and Accuracy in Other Tasks: Extension of the Binning Procedure

Task-switching procedures are not the only ones that rely upon cost or switch scores in RT as the dependent variable. Numerous other executive functioning tasks are designed in this manner and thus may share many of the same issues as task-switching paradigms. Just considering other attention control tasks (sometimes referred to as *inhibition*), the Stroop, flanker, Simon, and the three components of the attention network task (ANT; Fan, McCandless, Sommer, Raz, & Posner, 2002) are measured with RT-based difference scores. Specifically, in the arrow flanker task, subjects are asked to indicate which direction an arrow in the center of the screen is pointing. On congruent trials, this arrow is pointing in the same direction as the distracting (i.e., flanking) arrows that appear on either side of the target arrow. On incongruent trials, the target arrow is pointing in the opposite direction of the distracting arrows. Performance on the incongruent trials reflects attentional control, whereas performance on congruent trials is automatic, only requiring basic perceptual and motor abilities. Therefore, the dependent variable is an RT-based difference score of incongruent

Table 5. Correlations Among Composite Task-Switching Performance and Working Memory Tasks From the Reanalysis of Oberauer, Süß, Wilhelm, and Wittman (2003)

	Latency switch cost	Bin score
Reading span	.14	.29*
Computation span	.12	.21*
Spatial STM	.18*	.26*
Spatial coordination	.09	.24*
MU numerical	.22*	.29*
MU spatial	.17	.36*
WM composite	.20*	.37*

Note. $N = 131$. STM = short-term memory; MU = memory updating; WM = working memory. WM Composite is the composite score of the six other tasks in the table; latency switch cost and bin scores are composite scores of the four task switching tasks. The latency switch costs here are local switch costs, and are not log-transformed as in Oberauer et al. Correlations were multiplied by −1 such that a positive correlation indicates that subjects who did well on one also tended to do well on the other.
*$p < .05$.

Table 6. Predicting Working Memory Capacity From Bin Scores and Latency Switch Costs in Oberauer, Süß, Wilhelm, and Wittman (2003)

	Predictor	*R*	Adjusted *R*-Square	*SE*	*p*
		Model 1			
Step 1	RT Cost	.209	.036	.719	.016*
Step 2	RT Cost + Bin Score	.414	.158	.672	< .001*
		Model 2			
Step 1	Bin Score	.383	.146	.679	< .001*
Step 2	Bin Score + RT Cost	.414	.158	.672	.052

Note. $n = 131$. RT Cost = Composite latency switch cost score on all four task switching tasks; bin score = composite bin score on all four task switching tasks. The outcome variable is the composite score of the six working memory tasks.
*$p < .05$.

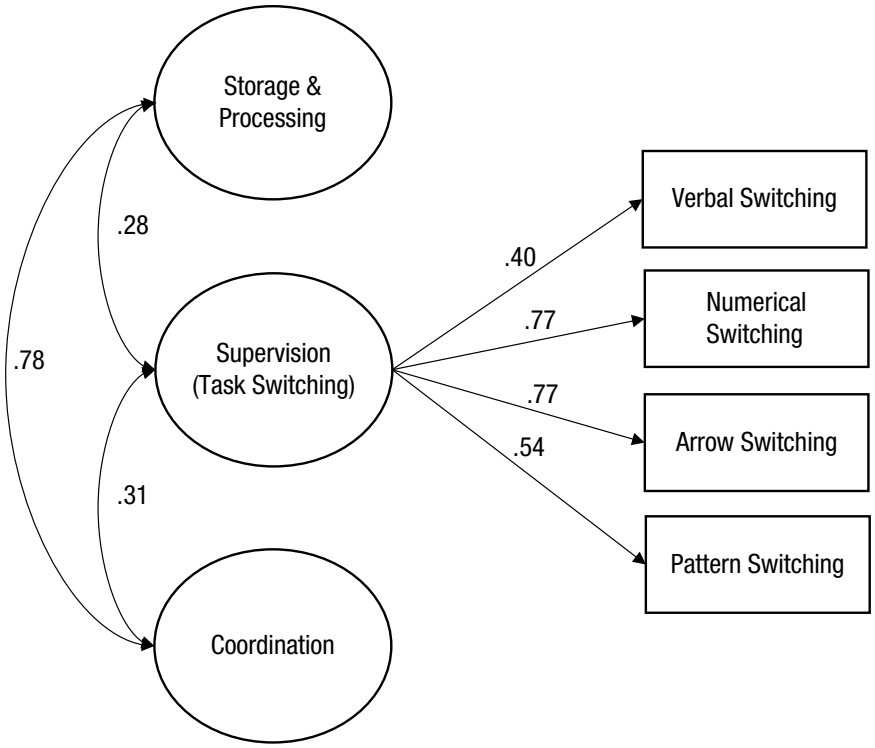

Fig. 4. Latent model from Oberauer, Süß, Wilhelm, and Wittman (2003) with log-transformed switch costs as the dependent variable for the task-switching tasks performed by 131 subjects. Correlations involving supervision were multiplied by −1 such that a positive correlation indicates better performance on both tasks. See Oberauer et al. (2013) for full description of tasks and executive functions. Figure adapted from "The Multiple Faces of Working Memory: Storage, Processing, Supervision, and Coordination," by K. Oberauer, H. M. Süß, O. Wilhelm, & W. W. Wittman, 2003, *Intelligence, 31*, p. 180.

and congruent trials, analogous to a latency switch cost. In the Stroop task, subjects see color words (e.g., "red") printed in different colored ink and are asked to name the color of the ink (as opposed to reading the word). Performance on trials in which the word is congruent with the color it is printed in is automatic, whereas responding to a trial in which the word is printed in a different color (e.g., the word "red" printed in green ink) requires inhibition and attention. The ANT has been touted as a potential diagnostic for attention-deficit problems in children (see Doyle, Biederman, Seidman, Weber, & Faraone, 2000), so understanding the relation between accuracy and speed of responding on this task would seem essential before it can be used as a diagnostic instrument.

Although the binning procedure was introduced as an alternative way to score task-switching data, we argue it can also be applied to these attention tasks as well.[11] Shipstead et al. (2015) used the flanker and Stroop tasks; we re-analyzed these two tasks to assess whether the binning procedure can be extended to these types of tasks. In the re-analysis, we found that binning the flanker data resulted in a stronger correlation with WMC than that found with latency switch costs (rs = .25 and .17, respectively), with the difference between these two correlations being significant ($p < .05$). Binning the Stroop data did not result in a statistically different correlation compared with that obtained with latency switch costs (rs = .24 and .26, respectively). These data are shown on Table 7.

A closer inspection of the Stroop task reveals that mean accuracy rate was very high, and more important, there appeared to be almost no individual differences in accuracy. Table 8 shows descriptive statistics of mean accuracy rate on the switch (for task switching) and incongruent (for Stroop and flanker) trials and the correlation between this accuracy and WMC in each of the four tasks in Shipstead et al. (2015) that we analyzed with the binning procedure. Accuracy on the Stroop task was the highest of all of these tasks and also had the smallest variability. Given this ceiling effect, it is not surprising that neither accuracy rate on the incongruent trials nor the Stroop bin scores correlated strongly with WMC. Table 9 shows a breakdown of accuracy for high- versus low-WMC individuals. These factors were likely the reason that incorporating both RT and accuracy in

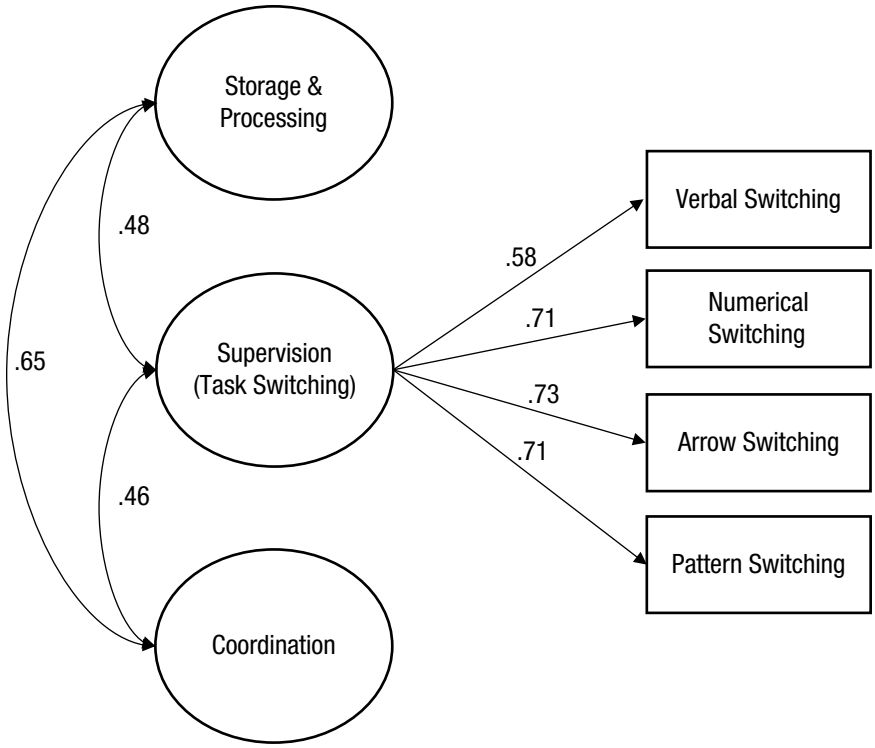

Fig. 5. Latent model from Oberauer, Süß, Wilhelm, and Wittman (2003) with the bin scores as the dependent variable for task-switching tasks performed by 131 subjects. Correlations involving supervision were multiplied by −1 such that a positive correlation indicates better performance on both tasks.

the Stroop task did not provide any significant differences in its relation to task switching. Simply put, if there are no differences in accuracy and accuracy is close to ceiling, it cannot provide any additional information. Alternatively, the Stroop may not be a good measure of executive functioning or the particular Stroop design used in Shipstead et al. (2015) may not have been difficult or demanding enough for individual differences to emerge.

It is worth noting that the tasks discussed so far are by no means a comprehensive list of tasks that utilize difference scores. Outside the realm of executive functions, the Implicit Association Test (IAT; Greenwald et al., 1998)is a salient example of a popular social psychology task that uses a difference score. This task putatively measures automatic (i.e., unconscious or implicit) associations related to stereotyping and discrimination. Despite its popularity, the IAT has also been criticized for (among other reasons) its psychometric integrity and lack of predictive and concurrent validity in terms of predicting real-world behaviors (see Landy, 2008). It is safe to say that the issue of RT-based difference scores is not confined to task-switching research or even the field of cognitive psychology. We return to this issue in later sections.

Gf and Task Switching: Theoretical Implications

Recently, a new framework for understanding the link between WMC and Gf has been proposed by our lab (Shipstead, Harrison, & Engle, 2014). In this framework, the strong relation between WMC and Gf occurs because the two constructs are driven by two separable, albeit highly related processes. Specifically, tasks that reflect WMC require retrieval and maintenance of goal-relevant

Table 7. Correlations Between Flanker and Stroop Tasks to Working Memory Capacity From Shipstead, Harrison, and Engle (2014)

	Flanker		Stroop	
	RT Cost	Bin	RT Cost	Bin
WMC	.17*[a]	.25*[a]	.26*[b]	.24*[b]
Difference	$p < .05$		$p > .05$	

Note. RT = reaction time; WMC = composite score on the four working memory tasks. Correlations were multiplied by −1, such that a positive correlation with WMC reflects better performance on the Stroop and flanker tasks for higher-WMC individuals.
[a]$n = 552$. [b]$n = 550$.
*$p < .05$.

Table 8. Correlation Between Switch/Incongruent Trial Accuracy and Working Memory Capacity From Shipstead, Harrison, and Engle (2014)

Task	DV	*N*	*M*	*SD*	Correlation to WMC
CatSwitch	Switch Trial Acc.	552	.84	.15	.50*
NumSwitch	Switch Trial Acc.	552	.82	.17	.45*
Flanker	Incongruent Trial Acc.	552	.92	.13	.25*
Stroop	Incongruent Trial Acc.	550	.93	.07	.11*

Note. DV = dependent variable; Acc = accuracy; WMC = composite score on the four working memory tasks; CatSwitch = category switch; NumSwitch = letter-number switch.

information, whereas tasks that reflect Gf require disengagement from outdated, now-irrelevant, information. Both of these processes are important in most tasks designed to measure either WMC or Gf: both processes help combat proactive interference, facilitate problem solving, and are hypothesized to be driven by executive control; for these reasons, a high correlation between Gf and WMC is usually observed.

A brief example illustrating this difference can be shown in matrix reasoning problems (e.g., Raven's advanced progressive matrices). In these tasks, the subject is shown a grid of figures (typically 3 × 3) with a piece missing and is tasked with choosing which of several possible figures belongs in the missing space. While matrix-reasoning tasks are often used as an indicator of Gf, performance is also highly correlated with WMC. According to our framework, WMC is required to form a stable representation of the problem in order to test hypotheses about potential rules governing the problem, whereas Gf is required to inhibit retrieval of previously tested and failed hypotheses that would otherwise interfere with solving the problem. That is, once a subject tests a particular hypothesis and recognizes that it is incorrect, retesting or becoming fixated on that particular hypothesis is detrimental. Thus, WMC and Gf are separable but also work in tandem, and an individual's WMC and Gf are both limited by his or her ability to control attention.

Table 9. Percentage of Accuracy on Switch/Incongruent Trials by Span in Shipstead et al. (2015)

	WMC	
Task	High	Low
CatSwitch[a]	93	73
NumSwitch[a]	92	71
Flanker[b]	95	88
Stroop[b]	94	92

Note. CatSwitch = category switch; NumSwitch = number-letter switch. Split by quartile of working memory capacity (WMC) composite score on the four working memory tasks.
[a]Percentage of accurate switch trials. [b]Percentage of accurate incongruent trials.

As discussed previously, reconfiguration theories posit that task switching depends largely on the task-set reconfiguration process. This reconfiguration likely involves both the activation of the new task set (forward-looking reconfiguration) and the inhibition of the previous, now-irrelevant task set (backward-looking reconfiguration), perhaps to differing degrees (Allport et al., 1994; Rogers & Monsell, 1995). Although these two processes cannot be disentangled in a standard task-switching experiment involving only two tasks, Mayr and Keele (2000) used a three-task paradigm to investigate which process is more important. Their findings supported the inhibition hypothesis (but see Lien, Ruthruff, & Kuhns, 2006), and they concluded that inhibiting the irrelevant task set has both facilitative and deleterious effects. For instance, the inhibition process permits successful and immediate task switching, but better inhibition also produces slowing if switching back to a formerly inhibited task set is necessary. In terms of our proposed dissociation between WMC and Gf, the inhibition of the previously relevant task set is a form of disengagement that should rely heavily on Gf. Thus, task switching should be more dependent on Gf than on WMC if inhibition of previous task sets is critical in switching procedures. Additionally, we would also expect high-Gf individuals potentially to be slower than low-Gf individuals on switch trials because their better ability to inhibit the task set also would result in a difficulty in reactivation of that task set on subsequent trials. In other words, being better able to inhibit (disengage) irrelevant information is beneficial in most situations, but if the information then becomes relevant again, it takes more time and processing for it to be re-activated.

In fact, that is precisely what was found in Shipstead et al. (2015). Recall that in this study, larger-WMC and Gf individuals were slower on the switch trials, but the binning procedure reanalysis produced a strong and positive relation between these two constructs and task switching. Additionally, as Table 10 shows, Gf predicted 7.3% unique variance in task-switching performance above and beyond WMC, whereas WMC predicted a negligible 2.2% unique variance to task-switching above and beyond Gf. These results suggest that although both WMC (maintenance) and Gf (disengagement) are good predictors of task-switching performance, Gf is more predictive, at least for cueing paradigms. Thus, this finding provides support that disengagement is a critical mechanism behind Gf and that inhibition of previous task sets is a more important process in switching between tasks.

Table 10. Predicting Task Switching Composite Scores From WMC and Gf in Shipstead, Harrison, and Engle (2014)

	Predictor	*R*	Adjusted *R*-Square	*SE*	*p*
			Model 1		
Step 1	WMC	.489	.238	.751	< .001
Step 2	WMC + Gf	.560	.314	.714	< .001
			Model 2		
Step 1	Gf	.540	.292	.724	< .001
Step 2	Gf + WMC	.560	.314	.714	< .001

Note. In the first model, task switching is predicted from working memory capacity (WMC) in Step 1 and then from both WMC and fluid intelligence (Gf) in Step 2. In the second model, the process is reversed.

Interim Summary

The binning procedure provides a substantial improvement in both reliability and validity over traditional difference scores. The procedure was used to analyze data from two different samples (Oberauer et al., 2003; Shipstead et al., 2015), two different types of task-switching tasks (alternating runs and cueing), and tasks of other executive functions (flanker and Stroop). It showed a consistent improvement in validity over switch costs in all cases in which accuracy differences were present. The improvements were evident with the zero-order correlations between tasks and also manifest at the latent level, as demonstrated in Oberauer et al. Viewed in total, the results from our analyses using the binning procedure provide strong evidence that task switching and WMC are highly related, a finding that has frequently eluded researchers. Additionally, the reanalysis provided support for our recently proposed view that different mechanisms are responsible for the latent constructs of WMC and Gf, with maintenance being responsible for WMC effects and disengagement responsible for Gf effects. It also supports the findings of Mayr and Keele (2000) that inhibition of previously active, but now-irrelevant, task sets is a crucial component of the task-set reconfiguration process. The following sections are dedicated to potential limitations and unaddressed questions regarding the binning procedure, how our findings can influence other areas of psychology, and what conclusions we draw from our results.

Potential Issues and Limitations of the Binning Procedure

In this section, we address further considerations of the binning procedure, along with its limitations. Specifically, we discuss (a) whether there is a justifiable reason for how inaccurate trials are scored, (b) whether it is problematic that difference scores are still part of the calculation of bin scores, (c) whether the binning procedure actually measures task-specific performance, and (d) in which circumstances the binning procedure is appropriate.

How accuracy is scored

One concern of the binning procedure is the manner in which inaccurate trials are scored (i.e., how much accuracy is weighted). Assigning a value of 20 to each inaccurate switch trial (a value twice as high as that assigned to the slowest accurate switch trial responses) is admittedly arbitrary. However, this number is in a justified range given normal task instructions. Subjects usually are instructed to be as quick and accurate as possible or, alternatively, to be as quick as possible without making errors. Thus, a reduction in accuracy should be penalized more than a reduction in speed. To test whether scoring inaccurate trials as a 20 perhaps artificially produced the large correlations between task switching and WMC, we analyzed data from Shipstead et al. (2015) using different penalties for inaccurate trials. Figure 6 shows the correlation between WMC and the binning scores with different penalties for inaccurate trials. The results show that using 20, at least for these data, does not lead to the strongest correlations between task switching and WMC, as higher penalties result in higher correlations between these two constructs.[12] Hence, there is nothing special about using the value of 20 as the punishment for inaccurate switch trials; a value of 15 or 50 would lead to similar conclusions, with only moderate differences in the strength of the correlation. This finding shows that the practice of taking accuracy into consideration is more important than the precise weight given to accuracy. On the other hand, given that the correlation to WMC does increase as switch trial accuracy is given more weight, we also concede that correlating accuracy rates on switch trials to WMC (or another criterion variable) is a good quick check to see if the binning procedure or any other method of incorporating RT and accuracy may lead to different results than switch costs. Another alternative is for researchers to use the binning procedure in a similar manner as Figure 6, testing multiple penalties to see what type of trend is produced. This practice likely provides more information than using a single static weight.

Reliance upon difference score

A second potential concern of the binning procedure is that a difference score still is used is part of the calculation. This concern is a function of the manner in which current task-switching paradigms are constructed. The paradigms are designed so that either a local or global switch cost can be calculated as the dependent variable,

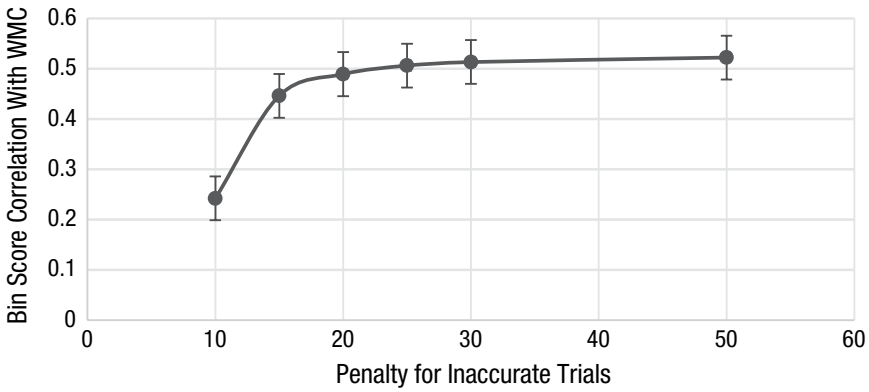

Fig. 6. Correlation between working memory capacity and bin scores when different penalties for inaccurate trials are used in data from Shipstead, Harrison, and Engle (2014). Error bars represent standard error of the mean. The penalties tested were 10, 15, 20, 25, 30, and 50. WMC = composite score on the four working memory tasks. The binning scores used in the correlation are composite scores from both task-switching tasks.

thus deriving a score that does not depend on switch costs is unavoidable. Within these types of paradigms, the best that can be achieved is a score that relies less on switch costs than others, which is something that is accomplished via the binning procedure.[13] Furthermore, there is a crucial distinction to make between how switch costs are normally calculated and how they are calculated in the binning procedure. The traditional switch cost score is a difference between two means, which leads to the correlation between the two component scores being subtracted out, and, thus, a larger portion of the leftover variance is error variance. In the binning procedure, the mean RT on nonswitch trials is subtracted from each and every individual RT on accurate switch trials. Because there is more variation among individual trials (as opposed to a mean, which is a constant), these scores are not as highly correlated. As a result, while difference scores are still utilized in the binning procedure, unreliability is not as much of a concern.

Construct-irrelevant variance

An additional potential criticism is that the binning procedure does not take baseline accuracy (e.g., accuracy on nonswitch or incongruent trials) into account and thus may capture variance that is reliable but not necessarily associated with the construct being measured.[14] That is, subjects making more overall errors in the task will have a worse bin score than subjects making fewer errors, but these errors could be the result of, say, differing processing efficiency and not due to a deficit in actual task-switching ability. Ostensibly, only counting incremental inaccurate switch trials for each subject (e.g., if a subject makes five errors on nonswitch trials and eight on switch trials, then he or she would be penalized for three inaccurate trials) could fix this issue. However, this is exactly how accuracy switch costs are calculated, which leads back to the original problem of using switch costs. Furthermore, accuracy switch costs are often entirely unreliable, even more so than latency costs. We devote the following paragraphs to discussing this concern, first by providing two pieces of evidence to counter this claim.

The first piece of evidence is that if the binning procedure artificially produced high correlations, or, alternatively, tapped into variance common to executive function tasks but not necessarily specific to task-switching ability, then we would expect other executive function tasks that were analyzed via binning to share this variance. To test this possibility, we performed a stepwise regression from the Shipstead et al. (2015) data using scores on the task-switching tasks (both latency switch costs and bin scores) and also the flanker task to predict WMC. The binning procedure takes two pieces of information into account: latency switch costs and accuracy on switch trials. Thus, if the binning procedure only captures variance common to executive functions (i.e., not specific to task switching), then the task-switching bin scores should not predict any unique variance in WMC beyond the latency switch costs (of the same task) and the bin scores of a different task of executive functioning. That is, if the correlation between WMC and task-switching bin scores is due to the binning procedure measuring processing efficiency or processing speed, then the same should be true of bin scores for the flanker. As the stepwise regression (Table 11) shows, however, this is not the case. The bin scores of the two task-switching tasks in Shipstead et al. predict a significant amount of unique variance in WMC above both the latency switch cost measures of these tasks and the bin scores of the flanker task. Furthermore,

the amount of unique variance in WMC and the task switching bin scores is quite high—16.3% in this sample. This provides support that the binning procedure measures an important aspect of task-switching ability that is common to WMC but not common to latency switch costs, flanker performance, or the binning procedure itself.

The second piece of evidence comes from data from Oberauer et al. (2003), who included eight measures of processing designed to measure processing efficiency separate from memory storage. If the increased correlation between task switching (when analyzed via binning) and WMC is due to the binning procedure reflecting processing efficiency, then the bin scores should not predict any unique variance in WMC beyond processing ability and latency switch costs. The stepwise regression in Table 12 shows that the binned task-switching data do indeed contribute significant unique variance to WMC beyond latency switch costs and processing ability. Although latency switch costs and processing ability predict a large and significant amount of variance in WMC (mostly processing, which explains nearly 26% of the variance in WMC), adding the bin scores to the model produces an additional 7.3% explained variance in WMC. These two regression models show that the increased magnitude of the correlations observed between WMC and task switching is not a product of the binning procedure measuring processing efficiency or another component common to executive functioning tasks; instead, it measures a component of task-switching ability that is both reliable and specific to task switching and WMC.

In a similar vein, one reviewer expressed concern that the binning procedure might penalize more error-prone or lower-ability individuals independent of their actual task-switching ability and suggested a diagnostic analysis of the binning procedure to test this concern. The question here is whether the relationship between WMC and task switching is still strong when nonswitch trial errors are penalized instead of switch trial errors.[15] If so, it could be problematic for the conclusions that we have made so far because it suggests that bin scores might reflect variance that is common to both nonswitch and switch trials and thus variance not related to task-switching processes. However, if the relation between WMC and the bin scores is weak when nonswitch errors are penalized, then we more confidently argue that binning method is indeed capturing systematic variance unique to task-switching mechanisms. Therefore, this analysis would show if performance on switch trials (in which task switching mechanisms are required) predicts WMC beyond accuracy on nonswitch trials (in which task-switching mechanisms are not utilized, and thus, performance differences cannot be attributed to task switching).

Table 11. Predicting Working Memory Capacity From Latency Switch Costs, Flanker Bin Scores, and Task-Switching Bin Scores From Shipstead, Harrison, and Engle (2014)

	Predictor	*R*	Adjusted R^2	*SE*	*p*
Step 1	RT cost + Flanker bin score	.332	.107	.799	< .001
Step 2	Task-switching bin score	.523	.270	.722	< .001

Note. RT (reaction time) cost = composite latency switch cost of category switch and letter-number switch; task-switching bin score = composite bin scores of category switch and letter-number switch.

We conducted this analysis on both the Shipstead et al. (2015) and Oberauer et al. (2003) data sets and found mixed results. In the Shipstead et al. data set, the correlation between bin score and WMC was still strong when nonswitch errors were penalized (r = .51 at the composite level) instead of switch errors and not statistically different from the relation between WMC and the normal bin scores (r = .49). Additionally, the correlation between the bin scores when nonswitch errors were penalized versus when switch errors were penalized was very strong (r = .91, p < .05), indicating that it did not matter whether nonswitch or switch trial errors were penalized. However, in the Oberauer et al. data set, penalizing nonswitch errors did not significantly predict WMC in two of the four task switching tasks (rs = .01 and .10, both p > .05) and in the other two predicted WMC in the opposite direction (rs = −.21 and .34, both p < .05). At the composite level, the correlation with WMC was in the opposite direction (r = −.27, p < .05). Furthermore, bin scores when switch trial errors were penalized and bin scores when nonswitch trial errors were penalized were not significantly correlated (r = −.17, p > .05).

Table 12. Predicting Working Memory Capacity From Processing Task Performance, Latency Switch Costs, and Task-Switching Bin Scores From Oberauer, Süß, Wilhelm, and Wittman (2003)

	Predictor	*R*	Adjusted R^2	*SE*	*p*
Step 1	Processing + RT costs	.524	.264	.632	< .001
Step 2	Task-switching bin score	.593	.337	.604	< .001

Note. Processing = composite performance on the eight processing tasks; RT (reaction time) costs = latency switch costs on the four task-switching tasks; task-switching bin score = composite bin scores of the four task-switching tasks. The criterion, working memory, is the composite score from the six working memory tasks.

What are the implications of these analyses? In the Oberauer et al. data set, penalizing nonswitch errors instead of switch errors did not result in a positive relation between task switching and WMC. This result supports our claim that the binning procedure reflects variance uniquely associated with the task-set reconfiguration process. This finding highly suggests that the binning procedure is a more valid and appropriate analysis than latency switch costs for this data set, likely because the binning procedure takes accuracy into account and is more reliable than switch costs.

In the Shipstead et al. data set, the results were not so clean. There was little difference in bin scores when nonswitch trial errors were penalized versus when switch trial errors were penalized. This finding potentially nullifies our previously discussed results and indicates that the binning procedure may be capturing variance not necessarily associated with the task-set reconfiguration process and instead results in an artificially high correlation between task switching and WMC.

We argue that this is not the case and provide three pieces of evidence to support our claim that the correlation to task switching and WMC in Shipstead et al. is not artificial. First, bin scores from the task-switching tasks predict unique variance in WMC beyond bin scores from other tasks, as previously discussed (see Table 11). Second, even nonswitch trials in Shipstead et al. (2015) involve switch costs because nonswitch trials and switch trials were in the same block of trials. Therefore, subjects would have been slower and more error prone on these nonswitch trials than they would have been if the nonswitch trials were in a pure-block design (or global switch costs). Thus, if subjects with smaller WMC showed more global switch costs than those with larger WMC, we would still expect to find a relation between WMC and bin scores when nonswitch trial inaccuracy was penalized (which we did).[16] Third, the task-switching tasks in Shipstead et al. (2015) had a demanding response deadline that resulted in increased errors, and it is quite likely that some subjects in this study were making a strategic speed-accuracy adjustment in response to being in the highly demanding task. If this adjustment is responsible for the high correlation between task switching and WMC, then it suggests there were individual differences in how subjects altered their performance in response to being in the demanding task-switching setting.

To test this hypothesis, we examined how subjects of differing WMC behaved after making errors in Shipstead et al. We found a strong and positive correlation between WMC and accuracy on all trials both preceding ($r = .52$, $p < .05$) and following ($r = .52$, $p < .05$) an error, meaning that high-WMC individuals were more accurate in general. What is more interesting is how individuals adjusted their RT after making an error. There was a weak correlation between WMC and RT on trials preceding an error ($r = .10$, $p < .05$) but a much larger correlation between WMC and RT after an error ($r = .27$, $p < .05$). The difference between these two correlations was statistically significant at the .05 level, which is direct evidence that subjects with larger WMC slowed down more after an error than subjects with smaller WMC. As a result, larger-WMC individuals were less likely to make an error on subsequent trials. This post-error slowing is likely the reason that higher-ability subjects were more accurate in general and is partially responsible for a strong correlation between task switching and WMC in Shipstead et al.

Taken together, these results suggest that the correlation between WMC and task switching that we have reported in the Shipstead et al. (2015) and Oberauer et al. (2003) data sets emerged for different reasons. In Oberauer et al., the binning procedure resulted in a high correlation between task switching and WMC because of its higher reliability than switch costs and because of accuracy being taken into account. We can confidently say that the bin scores for this data set capture variance associated with the task-set reconfiguration. In Shipstead et al., the binning procedure resulted in a high correlation between WMC and task switching two reasons: high-ability subjects made a speed-accuracy adjustment in response to a difficult and demanding task-switching setting and smaller-WMC subjects exhibited more global switch costs than larger-WMC subjects. The second point cannot be directly tested with the current data, but it is an alternative explanation for the strong correlation between WMC and bin scores when nonswitch errors are penalized instead of switch trial errors.

When should the binning procedure be used?

Finally, the conditions in which the binning procedure is appropriate should be considered. As stated previously, it can be applied to many non-task-switching tasks, so long as the task involves comparison of different trial types using a difference score and both RT and accuracy are meaningful. However, what particular qualities of the sample or task is the binning procedure best suited to assess? Both of the data sets that we reanalyzed in the present article were correlational studies with large samples, and in Shipstead et al. (the study with the more discrepant results between switch costs and bin scores), the sample was not only very large but also very diverse. It is possible that the binning procedure is best suited to these types of studies because it has a component in which trials are rank-ordered across all subjects. As such, it may be sensitive to nonnormal distributions such that the bin scores likely will mirror the ability level distribution of the sample[17] and require a larger sample size than

typical experiments. Also, all of the switching tasks analyzed in this article had 98 trials, which is common in task-switching research. However, given that two sources of information (RT and accuracy) are incorporated into one and that bin scores are not mere differences between means, it is likely that the binning procedure is more sensitive to the number of trials and requires more trials in order to be maximally reliable and valid. Last, as discussed in the previous section, individual differences in accuracy are necessary for the binning procedure to differentiate subjects better than analysis of switch costs alone. In Shipstead et al., the response deadline in the switching tasks drove overall accuracy rates down, likely facilitating the emergence of individual differences in accuracy rates. In contrast, accuracy in the Stroop task was very close to ceiling, and both high- and low-WMC individuals did not exhibit differences in mean accuracy rates. It is no surprise, then, that binning the Stroop data did not produce different results.

General Conclusions

WMC and task switching

Numerous studies have failed to find that working memory is important in an individual's ability to switch tasks. Our results suggest that methodological issues played a large role in these findings and that these data should be analyzed in more detail. For example, it is likely that a reanalysis of the data from Miyake et al. (2000) would yield results similar those of Shipstead et al. (2015) and Oberauer et al. (2003). A more direct comparison can even be done between the studies of Miyake et al. and Hughes et al. (2014) because both contained an antisaccade task and modern task-switching tasks. The correlations (*r*s) in Miyake et al. between these two were .17 (for the alternating-runs task) and .11 (for the local-global task), both nonsignificant ($\alpha = .05$). In Experiment 1 of Hughes et al., the correlation between their cued-switching task and the antisaccade was .185 using latency switch costs (in line with Miyake et al.) but .253 using bin scores. The same pattern of results thus would be expected of a reanalysis of Miyake et al. It is worth noting, however, that, of all the tasks in Miyake et al., their switching tasks seemed to correlate strongest with the Wisconsin Card Sorting Task (WCST). Miyake et al. noted that this task has been suggested to reflect both shifting between task sets and inhibition of inappropriate responses. Because of this relation, Miyake et al. suggested that task switching may be a function of inhibiting previously active, now-irrelevant task sets. We largely agree with this conclusion. In line with Mayr and Keele (2000), we suggest that backward-looking reconfiguration is a major process behind task switching. As discussed previously, this conclusion would explain the finding that Gf and task switching are more strongly related than WMC and task switching.

At a higher level, our results also have implications for both society and the individual. At the societal level, because switching between tasks relies on limited resources in working memory, individual differences in task-switching ability are linked to attention and intelligence and predict performance across a wide range of jobs and other demanding or otherwise stressful situations. Having the best personnel in many of these jobs is important because failure has major consequences, not just economically but in terms of human life. For instance, a failure to switch from Task A to Task B and back to Task A for soldiers, doctors or surgeons, or air-traffic controllers could be both economically expensive and fatal to themselves or others. Knowing that limited attentional resources in working memory is linked to this ability can help employers select the best candidates to minimize potential losses and provide both employers and applied psychologists with the knowledge to create job situations in which task switching is facilitated, minimizing the potential for mistakes. At the individual level, task switching is a common, everyday behavior that people engage in often without realizing it. Our results suggest that in a stressful or cognitively demanding situation, the working memory load imposed likely hinders the ability to properly switch from one task to another. In these circumstances, it is better to focus on performing one action rather than on switching back and forth between multiple actions or attempting to minimize the cognitive load. Otherwise, either mistakes or slowing will occur. In some circumstances, slowing or errors are not particularly costly, but in others, for example, driving on a busy freeway or in a residential area just as grade school children are being sent home, they can be quite costly.

Combining RT and accuracy

Throughout this article, we have used the binning procedure to demonstrate how RT and accuracy can be combined into a single metric to overcome the concerns associated with latency switch costs and have also shown that this particular procedure can be used to measure non-task-switching tasks as well. Specifically, we have argued that WMC and task switching are highly related constructs (as are Gf and task switching), that inhibition of previously active task sets are crucial in task-set reconfiguration, and that consideration of both RT and accuracy in analysis can potentially be beneficial in a wide array of tasks. Furthermore, we have provided support for our theory that disengagement is a critical mechanism driving Gf.

However, we are not arguing that the binning procedure is the only or even the best method available for combining RT and accuracy. It is quite possible that the other reliable scoring technique presented in Hughes et al. (2014; the rate residual procedure) would perform just as well or even better, particularly with lower sample sizes. In addition, new scoring techniques or switching paradigms could arise that allow researchers to avoid using difference scores altogether. We hope that future research will be directed toward modeling task switching and different task-switching analyses (like the binning procedure), further validating existing methods of incorporating RT and accuracy, and developing new methods for this purpose. Furthermore, although we have provided additional evidence in support of the reliability and validity of the binning procedure, our larger goal is not to claim that it is a superior analysis but rather to make a point about larger issues that are relevant to many areas of psychology.

The main issue we would like to emphasize is that in many different psychological tasks, both RT and accuracy ought to be considered in the analysis for the conclusions to be valid, particularly when subjects are grouped on the basis of individual or developmental differences. Specifically, if both RT and accuracy represent important processes to the task at hand, then researchers should either measure both or, at a minimum, control for one. For example, Wickelgren (1977) argued strongly that researchers studying cognition should use methods that measure speed-accuracy trade-offs, claiming that speed-accuracy studies were vastly superior to RT studies and that conclusions drawn from typical RT studies could almost always be called into question. Using task switching as an example, we would emphasize that latency switch costs come with the assumptions that there will be no meaningful differences in accuracy across subjects and no individual differences in how much emphasis is placed on both speed and accuracy. Violations of these assumptions call into the question the validity of any conclusions made from these measures. These assumptions are frequently violated in cognitive tasks, particularly in studies of individual differences in which subjects have been preselected on the basis of WMC or developmental differences. In these circumstances, individual differences are inherent to different groups, and thus, we would expect there to be large differences in executive functioning across groups. Although these violations are most problematic in tasks that measure executive functioning with a difference score (like task switching and many attention control tasks), the violations are problematic in other tasks, such as the IAT. More generally, in any experimental work, researchers should exercise caution in using dependent variables that do not take into account both speed and accuracy or, alternatively, should employ paradigms that control for one of the two. Researchers should also consider the psychometric properties of their dependent variables.

Acknowledgments

We thank K. Oberauer for providing us with data that were used in this manuscript as well as his extensive comments. We thank M. Bunting for making the initial suggestion to perform some of the analyses in this manuscript and also for providing his extensive comments. We thank J. Foster, T. Harrison, M. Kane, and N. Cowan for their helpful comments on this manuscript.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

Part of this work was supported by Grants N00014-12-1-0406 and N00014-12-1-1011 from the Office of Naval Research to Randall Engle.

Notes

1. Switch costs for accuracy can also be calculated in an analogous manner, and researchers sometimes look at both accuracy and latency switch costs to see if the pattern of results is coherent.

2. See Baddeley (1992) for a brief historical review of working memory research. This review includes a description of the tripartite model of working memory involving the central executive, phonological loop, and visuospatial sketch pad.

3. We provide only a brief summary of this study and focus on only the parts relevant to this article. For a full description of the study, see Shipstead et al. (2015).

4. The true (raw) correlations between these variables were positive because task-switching tasks are designed such that a higher score indicates worse performance. However, throughout this article, we multiply correlations involving task switching and WMC or Gf by −1 for ease of comparison. Thus, a positive correlation indicates that individuals who did well on one task also tended to do well on the other, whereas a negative correlation indicates that individuals who performed well on one task tended to do poorly on the other. Any reference we make in the text takes this adjustment into consideration as well; therefore, in a statement such as "positive correlation between task switching and WMC," the reader should assume the −1 multiplication.

5. Miyake et al. (2000) argued that at the latent level, task switching is related to other executive functions (inhibition and updating). However, their zero-order correlation matrix suggested that task switching was weakly related to these other executive functions. Specifically, both their alternating-runs and local-global task switching tasks significantly correlated with only one of the other six tasks of executive functioning at the zero-order level ($\alpha = .05$), with many of the correlations being in the single digits. In Friedman et al. (2006), task switching

correlated with inhibition and updating in the range from .13 to .30 at the task level.

6. It has been argued that studies taking speed-accuracy trade-offs into consideration are vastly superior to traditional RT studies, such that conclusions made from RT studies can always be called into question (e.g., Luce, 1986; Wickelgren, 1977). For a review of speed-accuracy trade-offs, see Heitz (2014).

7. It should be noted that the extent to which the difference score paradox is a legitimate concern is highest contested. For example, Overall and Woodward (1975) showed that difference scores with a reliability of zero actually maximize the power of paired *t* test statistics, although they conceded that such difference scores are likely problematic for correlational research. Chiou and Spreng (1996) argued that the assumptions of equal variance, reliability, and relation to other variables among the component scores used to calculate the difference score rarely hold, and violations to these assumptions lead to a more reliable difference score. Tisak and Smith (1994) argued that difference scores are an acceptable dependent variable so long as the component scores are reliable and are not highly correlated.

8. Measures of internal consistency, particularly Cronbach's α, confound homogeneity of the items with the true reliability of the task. Thus, assessing the reliability of a highly homogeneous task, such as most task-switching procedures, result in an inflated reliability estimate.

9. The primary differences in the two task-switching tasks in Shipstead et al. (2015) were minor. They only differed in the type of judgment required and the stimuli used. Category switch had univalent word stimuli appear after a heart or a cross cue and asked for a living/nonliving or large/small judgment. Letter-number switch had bivalent letter-number pairs appear either below or above a horizontal line and asked for an even/odd or vowel/consonant judgment. Therefore, a correlation larger than .29 would be reasonably expected between these tasks.

10. In terms of Gf and task switching, the binning procedure revealed a similar relation as in WMC and task switching. That is, individuals who performed better on the Gf tasks also performed better on the task-switching tasks when the binning procedure was used to analyze the data. The relation between Gf and task switching was actually stronger than that of WMC and task switching.

11. The manner in which the binning procedure was designed makes it appropriate for analyzing any task of sequentially presented individual trials in which RT and accuracy are collected, provided there are two different trial types that can be compared and a theoretical reason to compare them. Such tasks are frequently analyzed with RT-based difference scores.

12. However, the correlation between task-switching accuracy and WMC when 50 was used as the penalty for inaccuracy was not significantly different than when 20 was used. It should also be noted that these results are from a data set in which the latency switch costs were negatively related to WMC; this particularity may be driving the results shown on Figure 6.

13. There are tasks that require task switching and do not rely on switch costs. However, these tasks are not designed to be pure measures of task switching as they confound switching ability with other executive control processes such as memory updating and WMC (e.g., Logan, 2004).

14. We thank both M. Kane and K. Oberauer for this observation.

15. Recall that in the binning procedure, inaccurate switch trials are scored as 20, but inaccurate nonswitch trials are not taken into account. Therefore, only errors in switching are penalized, but nonswitch errors are ignored. The reviewer's question was whether the pattern of results would be the same if the binning procedure penalized nonswitch errors instead of switch errors.

16. Unfortunately we cannot test this hypothesis directly with the data we have, because there were no pure-block conditions in Shipstead et al. that would allow us to calculate global switch costs.

17. In both Shipstead et al. (2015) and Oberauer et al. (2003), the bin scores were essentially normally distributed, with skewness and kurtosis under 1.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, *131*, 30–60.

Allport, D. A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umita & M. Moscovitch (Eds.), *Attention and performance XV* (pp. 421–452). Hillsdale, NJ: Erlbaum.

Altmann, E. M., & Gray, W. D. (2008). An integrated model of cognitive control in task switching. *Psychological Review*, *115*, 602–639.

Arrington, C. M., & Logan, G. D. (2004). The cost of a voluntary task switch. *Psychological Science*, *15*, 610–615.

Baddeley, A. D. (1992, January 31). Working memory. *Science*, *255*, 556–559.

Baddeley, A. D., Chincotta, D., & Adlam, A. (2001). Working memory and the control of action: Evidence from task switching. *Journal of Experimental Psychology: General*, *130*, 641–657.

Baddeley, A. D., Gathercole, S. E., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*, 158–173.

Case, R., Kurland, M. D., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology*, *33*, 386–404.

Chevalier, N., Sheffield, T. D., Nelson, J. M., Clark, C. A. C., Wiebe, S. A., & Espy, K. A. (2012). Underpinnings of the costs of flexibility in preschool children: The roles of inhibition and working memory. *Developmental Psychology*, *37*, 99–118.

Chiou, J., & Spreng, R. A. (1996). The reliability of difference scores: A re-examination. *Journal of Consumer Satisfaction, Dissatisfaction, and Complaining Behavior*, *9*, 158–167.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*, 769–786.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral & Brain Sciences*, *24*, 87–185.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"—or should we? *Psychological Bulletin*, *74*, 68–80.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in WM and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*, 422–433.

Doyle, A. E., Biederman, J., Seidman, L. J., Weber, W., & Faraone, S. V. (2000). Diagnostic efficiency of neuropsychological test scores for discriminating boys with and without attention deficit-hyperactivity disorder. *Journal of Consulting and Clinical Psychology*, *68*, 477–488.

Edwards, J. R. (2001). Ten difference score myths. *Organizational Research Methods*, *4*, 265–287.

Emerson, M. J., & Miyake, A. (2003). The role of inner speech in task switching: A dual-task investigation. *Journal of Memory and Language*, *48*, 148–168.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*, 19–23.

Engle, R. W., Carullo, J. J., & Collins, K. W. (1991). Individual differences in working memory for comprehension and following directions. *Journal of Educational Research*, *84*, 253–262.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*, 309–331.

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

Fan, J., McCandless, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, *14*, 340–347.

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*, 172–179.

Grant, D. A., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology*, *38*, 404–411.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York, NY: McGraw-Hill.

Hambrick, D. Z., Oswald, F. L., Darowski, E. S., Rench, T. A., & Brou, R. (2010). Predictors of multitasking performance in a synthetic work paradigm. *Applied Cognitive Psychology*, *24*, 1149–1167.

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working memory training may increase working memory capacity but not fluid intelligence. *Psychological Science*, *24*, 2409–2419.

Heitz, R. P. (2014, June11). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, Article 150. doi:10.3389/fnins.2014.00150

Hughes, M. M., Linck, J. A., Bowles, A. R., Koeth, J. T., & Bunting, M. F. (2014). Alternatives to switch-cost scoring in the task switching paradigm: Their reliability and increased validity. *Behavior Research Methods*, *46*, 702–721.

Jersild, A. T. (1927). Mental set and shift. *Archives of Psychology*, *14*, 5–81.

Kane, M. J., Bleckley, M. K., Conway, A. R., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General*, *130*, 169–183.

Kane, M. J., Conway, A. R. A., Hambrick, D. Z., & Engle, R. W. (2007). Variation in working memory capacity as variation in executive attention and control. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. Towse (Eds.), *Variation in working memory* (pp. 21–48). New York, NY: Oxford University Press.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*, 66–71.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhem, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology*, *133*, 189–217.

Kiesel, A., Stenhauer, 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.

Kiesel, A., Wendt, M., & Peters, A. (2007). Task switching: On the origin of response congruency effects. *Psychological Research/Psychologische Forschung*, *71*, 117–125.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity. *Intelligence*, *14*, 389–433.

Landy, F. J. (2008). Stereotypes, bias, and personnel decisions: Strange and stranger. *Industrial and Organizational Psychology*, *1*, 379–392.

Liefooghe, B., Barrouillet, P., Vandierendonck, A., & Camos, V. (2008). Working memory costs of task switching. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 478–494.

Liefooghe, B., Vandierendonck, A., Muyllaert, I., Verbruggen, F., & Vanneste, S. (2005). The phonological loop in task alternation and task repetition. *Memory*, *13*, 650–660.

Lien, M. C., Ruthruff, E., & Kuhns, D. (2006). On the difficulty of task switching: Assessing the role of task-set inhibition. *Psychonomic Bulletin & Review*, *13*, 530–535.

Logan, G. D. (2004). Working memory, task switching, and executive control in the task span procedure. *Journal of Experimental Psychology*, *133*, 218–236.

Logan, G. D., & Gordon, R. D. (2001). Executive control of visual attention in dual-task situations. *Psychological Review*, *108*, 393–434.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 22–38). Madison: University of Wisconsin Press.

Luce, D. (1986). *Response times: Their role in inferring elementary mental organization*. New York, NY: Oxford University Press.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*, 163–203.

Mayr, U., & Bell, T. (2006). On how to be unpredictable. *Psychological Science*, *17*, 774–780.

Mayr, U., & Keele, S. W. (2000). Changing internal constraints on action: The role of backwards inhibition. *Journal of Experimental Psychology: General*, *129*, 4–26.

Mayr, U., & Kliegl, R. (2000). Task-set switching and long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1124–1140.

Meiran, N. (1996). Reconfiguration processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1423–1442.

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 Sciences*, *7*, 134–140.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H. M. (2005). Working memory and intelligence—Their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*, 61–65.

Oberauer, K., Suß, H. M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity-facets of a cognitive ability construct. *Personality and Individual Differences*, *29*, 1017–1045.

Oberauer, K., Süß, H. M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence*, *31*, 167–193.

Oberaurer, K., Süß, H.M., Wilhelm, O., & Sander, N. (2007). Variation in working memory capacity as variation in executive attention and control. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 21–48). New York, NY: Oxford University Press.

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin*, *82*, 85–86.

Peter, J. P., Churchill, G. A., & Brown, T. J. (1993). Caution in the use of difference scores in consumer research. *Journal of Consumer Research*, *19*, 655–662.

Raven, J. C. (1941). Standardization of progressive matrices. *British Journal of Medical Psychology*, *19*, 137–150.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*, *28*, 164–171.

Rogers, R. D., & Monsell, S. (1995). The costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*, 207–231.

Rubin, O., & Meiran, N. (2005). On the origins of the task mixing cost in the cueing task-switching paradigm. *Journal of Experimental Psychology*, *31*, 1477–1491.

Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive control of cognitive processes in task switching. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 763797.

Shah, P., & Miyake, A. (1996). The separability of working memory resources from spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General*, *125*, 4–27.

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2014, November). *Working memory capacity and fluid intelligence: Maintenance and disengagement*. Paper presented at the 55th annual meeting of the Psychonomics Society, Long Beach, CA.

Shipstead, Z., Harrison, T. L., Trani, A. N., Redick, T. S., Sloan, P., Bunting, M. F., . . . Engle, R. W. (2015). *The unity and diversity of working memory capacity and executive functions: Their relationship to general fluid intelligence*. Manuscript submitted for review.

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

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251.

Stoet, G., & Snyder, L. H. (2007). Extensive practice does not eliminate human switch costs. *Cognitive, Affective, & Behavioral Neuroscience*, *7*, 192–197.

Tisak, J., & Smith, C. S. (1994). Defending and extending difference score methods. *Journal of Management*, *20*, 675–682.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*, 127–154.

Unsworth, N., & Engle, R. W. (2008). Speed and accuracy of accessing information in working memory: An individual differences investigation of focus switching. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 616–630.

Vandierendonck, A., Liefooghe, B., & Verbruggen, F. (2010). Task switching: Interplay of reconfiguration and interference control. *Psychological Bulletin*, *136*, 601–626.

Wickelgren, W. A. (1977). Speed-accuracy tradeoffs and information processing dynamics. *Acta Psychologia*, *41*, 67–85.