Reaction Time in Differential and Developmental Research: A Review and Commentary on the Problems and Alternatives

Christopher Draheim, Cody A. Mashburn, Jessie D. Martin, Randall W. Engle

Funded by:

- Office of Naval Research

Abstract

Reaction time is believed to be a good indicator of the speed and efficiency of mental processes and is a ubiquitous variable in the behavioral sciences. Despite this popularity, there are numerous issues associated with using reaction time, specifically in differential and developmental research. Here, we identify and focus on two main problems – unreliability and sensitivity to speed-accuracy interactions. The use of difference scores is a primary factor that leads to many reaction time measures having demonstrably low reliability, and reaction time measures in general often do not properly account for speed-accuracy interactions. Both factors jeopardize the validity and interpretability of results based on reaction time. We evaluate conceptually and empirically how these issues affect individual differences research. Although the empirical evidence we provide are primarily within the domains of attention control and task switching, we highlight examples from various other areas of psychological inquiry. We also discuss many of the statistical and methodological alternatives available to researchers conducting correlational studies. Ultimately, we encourage researchers comparing individuals of differing cognitive and developmental levels to strongly consider using these alternatives in lieu of reaction time, specifically reaction time difference scores.

Keywords*:* reaction time, individual differences, difference scores, reliability, speed-accuracy tradeoff, attention control, task switching, Implicit Association Test

Public Significance Statement: This review identifies and discusses the problems with the use of reaction time, particularly reaction time differences, in assessing how individuals differ from one another. Given these problems, a variety of conclusions and theoretical accounts stemming from reaction times in individual differences studies may be misinformed. Examples

include the efficacy of some clinical techniques, the measurement of racial bias, and the measurement of attention.

Reaction time in differential and developmental research: A review and commentary on the problems and alternatives

The distinction between differential and experimental psychology is often underappreciated among behavioral researchers (see Cronbach, 1957). Differential research is mainly correlation-based and investigates how people differ from one another (individual differences). In contrast, experimental research is primarily ANOVA-based and investigates how the performance of groups or conditions vary systematically, and causally, due to experimental manipulation. Historically, these two approaches have been interested in many of the same questions, but separately so, creating a clear divide as well as confusion between researchers following the two different approaches.

Distinguishing between experimental and differential approaches is important for a variety of reasons. Foremost, the two approaches have different concerns regarding the reliability of their measures. An experimental effect is reliable to the extent that it consistently replicates across studies and labs. However, a measure is reliable for differential purposes to the degree that it consistently rank-orders individuals across measurements (c.f., Hedge, Powell, & Sumner, 2018). A helpful way to conceptualize this is that experimental researchers are interested in maximizing within-subjects variance, whereas differential researchers are interested in maximizing between-subjects variance. [1] Because of this, measures can be reliable for one

[1] This characterization is true but somewhat misleading because experimental designs can involve either within-subjects contrasts, between-subjects contrasts, or both. The difference is that experiments treat subjects as being equal. Even though experiments can be between-subjects designs with subjects receiving different treatments, random assignment is used to mitigate the effects of between-subject differences by distributing between-subjects variance in each group. This problem is avoided in within-subjects designs. But, in differential research, subjects differing from one another is essential in maximizing variance and the strength of correlations. As such, experimental psychologists seek to minimize between-subject variance to the extent possible, whereas differential researchers instead rely on this variance.

approach but not the other. For an experimentalist, having thirty identical subjects would be perfectly fine, even desirable, if a robust effect emerged, whereas any researcher interested in individual and developmental differences would find those same data unusable.

It is not intuitive that performance on a well-established experimental paradigm can reliably produce robust experimental effects and yet fail to produce reliable and valid individual differences. This phenomenon is becoming increasingly acknowledged in the literature (e.g., Hedge et al., 2018; Hughes, Linck, Bowles, Koeth, & Bunting, 2014; Logie, Della Sala, Laiacona, Chalmers, & Wynn, 1996; Fisher, Medaglia, & Jeronimus, 2018; Paap & Sawi, 2016; Rey-Mermet, Gade, & Oberauer; 2018; Ross, Richler, & Gauthier, 2015; Rouder & Haaf, 2018; Whitehead, Brewer, & Blais, 2018). The findings of Logie, et al. (1996) nicely illustrate the difference between experimental and differential approaches. They investigated the generalizability of auditory and visual word length and phonological similarity in verbal short-term. These effects are robust in the literature, and they unsurprisingly found strong aggregate (group mean) effects**.** However, in their first experiment (N = 251), only 57% of their subjects showed all four effects of interest at the individual level. Further, they retested 40 subjects in a second experiment and found no statistically significant correlation between the initial test and retest for three of the four effects (r < .10), and only a moderate correlation for the fourth effect (r = .31). Meaning that whether or not a subject showed the effect in the initial test was not predictive of that individual showing the effect in the retest session. Crucially, the overall effects of interest still emerged at the aggregate level, despite the poor test-retest reliability of the individual scores. This is an important finding because it demonstrates how experimental effects can be reliable at the group level, but at the individual level they may not be, resulting in scores that provide little information about that individual's cognition.

Hedge, Powell, and Sumner (2018) reasoned that popular experimental tasks likely became popular for the very reason that makes them ill-suited to individual differences – the minimization of between-subject variability. They highlighted several paradigms in cognitive psychology impacted by the experimental vs. differential distinction, some of which we discuss in more detail in later sections. Along these lines, Rouder and Haaf (2018) noted that there are numerous reasons for individual differences researchers to be confident in using established experimental tasks, such as the repeated demonstrations of their robustness and internal validity, and yet performance on these tasks often does not correlate as expected.

Perhaps the hallmark example of a task being good in one context but not the other is the color Stroop task (MacLeod, 1991; Stroop, 1935). The color Stroop is a historic and useful task for experimental researchers and the Stroop effect is one of the most reliable effects in psychology. But, as noted by several researchers (e.g., Hedge et al., 2018; Paap & Sawi, 2016; Rouder & Haaf, 2018), correlations involving Stroop performance are highly attenuated, and Stroop tasks seem particularly poor for assessing individual differences.[2]

In this article, we more thoroughly discuss the psychometric and measurement problems related to reaction time. We offer conceptual arguments, supported by empirical evidence, to explain why reaction time, and subtraction methodology (difference scores) in particular, are problematic ***in differential and developmental contexts*****.** After this discussion we outline numerous alternatives that are available to researchers. We ultimately conclude that researchers

[2] It is important for the reader to keep in mind that reliability and validity are not inherent properties of a task, but rather measures have a certain reliability and validity for a specific purpose and in that context (e.g., Streiner & Norman, 1995). When we discuss some of the reliability and validity concerns regarding many common reaction time measures, the reader should not interpret this as an indictment of the task or paradigm as a whole, but as a criticism of that specific score and for that particular usage. For example, our concerns about the color Stroop relate to research using interference effects for individual differences or differential purposes, but we do not question the utility of Stroop tasks for experimental purposes.

interested in individual and developmental differences ought to strongly consider using one of these alternatives instead of pure reaction time or reaction time difference scores.

What constitutes acceptable reliability?

There are multiple factors that determine whether a measure is adequately reliable. These include the type of reliability estimate used (see Carmines & Zeller, 1979)[3], the reliability of alternative measures, the ability, process, or attitude that is being measured, the population of interest, the research question, and how the test scores are to be used. Nunnally (1964) argued that .70 is a *rough* minimum for exploratory research, but he recommended higher thresholds for other purposes: .80 for basic research interested in correlations or experimental effects; .90 for applied research; and .95 for high stakes testing (testing with consequences for the test-taker).[4] A cursory Google search suggested that these values are in line with other recommendations, with most sources citing .80 as the minimum threshold for good reliability of an ability test and .70 for a personality test (e.g., DeVellis, 1991). The takeaway here is not necessarily the specific values, but that the intended application of the test along with the reliability of available alternatives is critically important in assessing reliability. Forays into new areas of research can be successful with less reliable measures, whereas established fields of research should have a higher standard, and instances of job selection, job placement, legal matters, clinical diagnoses, college admissions, etc. ought to have the highest standards of all because of the consequences

[3] Test-retest reliability and internal consistency are the types of reliability relevant for our discussion. Test-retest reliability is assessed by correlating performance on the same measure at two (or more) separate times. Internal consistency involves correlating performance on individual items with each other item (Cronbach's Alpha; Cronbach, 1951) or correlating performance on half of the items with performance on the other half (e.g., even-odd or first and second halves of the test), known as split-half. Test-retest procedures result in lower estimates because numerous sources of construct-irrelevant variance can be introduced, such as reactivity or the respondent being in a different emotional or cognitive state during the separate administrations. The ability or construct in question may even change, meaning the true score is different for each administration. On the other hand, internal consistency procedures produce inflated scores to the extent that test items are similar to one another, and, because the test is only administered once, state-dependent factors cannot be teased apart, as is possible with test-retest.

[4] Lance, Butts, and Michels (2006) argued that this .70 recommendation for exploratory research has been taken out of context and improperly cited as the threshold for acceptable reliability in general.

for the test-taker. Further, the construct in question also requires consideration as to whether a measure has acceptable reliability or not.

What constitutes acceptable reliability is therefore situational and not straightforward. But, for purposes of discussing basic individual differences research, we regard reliability estimates below .70 as problematic, measures in the .70s as borderline, and measures at or above .80 as closer to the ideal.[5] Although we use Nunnally's (1964) standards as a guideline to help frame our discussion of reliability, we emphasize that researchers must decide for themselves what values are appropriate for their needs (c.f., Trafimow, 2015).

What are difference scores?

Difference scores follow the subtraction method tradition of Hermann von Helmholtz and Donders (1868/1969) – a subject's performance in one condition is subtracted from their performance in another. They come in different forms and, based on their form and application, are variously known as change scores, gain scores, residualized scores, cost effects, congruence effects, discrepancy effects, conflict effects, and/or interference effects.[6] For instance, a gain score is the difference of performance on the same test at two different time points, and is often used to assess an individual's improvement from treatment or training. Whereas an interference or conflict effect is the difference of performance on two different trial types of the same task, typically with minimal temporal separation, and is an indicator of the degree to which an individual is affected by cognitive interference. An example would be congruent and incongruent

[5] For reference, most of our accuracy-based measures of working memory capacity, fluid intelligence, and attention control produce scores with internal consistencies in the .80 - .90 range (see Table 1 in the Appendix), with the complex span task partial scores being .77 - .83 in terms of test-retest reliability (see Redick et al., 2012). The lower reliability estimates for absolute span scores (.63 - .83 for internal consistency and .62 - .77 for test-retest reliability across two studies) is one of the reasons that we have avoided assessing working memory capacity with absolute span scores in recent years.

[6] This is not an exhaustive list. Furthermore, nomenclature for types of difference scores is not always consistent; some researchers may refer to the same type of difference score by different names, or use the same label for two conceptually different types of difference scores.

trials of a Stroop task. Regardless of their form, difference scores all involve calculating a difference between a baseline measure and a related measure of interest.

Difference scores are ubiquitous in the behavioral sciences

The appeal of using difference scores is clear. Early in their careers, psychologists learn the importance of properly controlling for irrelevant and error variance. Experimental psychologists isolate variance of interest through means such as random assignment, active control groups, and effective experimental manipulations. Difference scores are ostensibly a simple and effective method of controlling irrelevant variance and isolating effects of interest, and thus seem an improvement over a simple mean reaction time.

Difference scores are therefore used in numerous areas of scientific inquiry and for a wide variety of research purposes. In cognitive psychology, difference scores are common in the measurement of executive functioning and frequently produce robust experimental effects. Notable examples include the Stroop (Stroop, 1935), Simon (Simon & Rudell, 1967), and flanker interference effects (Eriksen & Eriksen, 1974), as well as switch costs in task-switching paradigms (Monsell, 2003), which all contrast within-subject performance in two different but highly similar conditions. Difference scores are common in assessing cognitive control, for example with post-error slowing, which is the tendency for an individual to respond more slowly on trials immediately following an error trial (e.g., Dutilh et al., 2011; Rabbitt, 1966) and sequential effects (how trial order affects performance; e.g., Whitehead, Brewer, & Blais, 2018). The Attention Network Test (Fan, McCandliss, Sommer, Raz, & Posner, 2002) produces three difference scores purported to reflect three different components of attention, and is widely used by cognitive neuroscientists and developmental researchers (e.g., Callejas, Lupiáñez, Funes, & Tudela, 2005; Fuentes & Campoy, 2008; Konrad et al., 2005). Other areas of cognitive

psychology which use difference scores include the assessment of face recognition and processing (DeGutis, Wilmer, Mercado, & Cohan, 2013; Ross, Richler, & Gauthier, 2015), assessment of semantic priming effects (e.g., Sperber, McCauley, Ragain, & Weil, 1979), and sequential learning (Urry, Burns, & Baetu, 2015). In psychiatry, body image dissatisfaction is measured with difference scores (e.g., Cafri, van der Berg, & Brannick, 2010). In clinical psychology, gain scores are among the most common ways to assess therapeutic impact and treatment efficacy (Steketee & Chambless, 1992; Gottman & Rushe, 1993). Gain scores are similarly common in social psychology (e.g., Collins, 1996), one example being the Implicit Association Test which uses difference scores to assess implicit bias (Greenwald, McGhee, & Schwartz, 1998). Organizational behavior researchers and behavioral economists (see Edwards, 1994; 2001; Johns, 1981) frequently employ gain and change scores, as do consumer researchers (see Peter, Churchill Jr., & Brown, 1993). Difference scores are used are commonly used in survey research (e.g., Kessler, 1977). Finally, longitudinal studies often use difference scores to track the increase or decrease of a variable of interest over time (see Rogosa, Brandt, & Zimowski, 1982). This is by no means an exhaustive list, as any investigator interested in measuring change is likely to consider using a difference score in some form.

Subtraction methodology has undoubtedly been an invaluable tool for researchers and has advanced the understanding of human behavior. Here, we do not argue against the use of difference scores entirely, but, rather, we express caution in their application specifically in differential and developmental research. This is especially true when tasks and paradigms that are highly reliable in experimental studies are used *without modification* in differential and/or developmental contexts, as investigators are likely to falsely assume that the same measures that produce robust experimental effects will also consistently rank-order subjects.

Conceptual arguments against difference scores

Understanding the problem with difference scores first requires a discussion about the calculation of their reliability. The true formula for assessing reliability of difference scores is quite long (see Cronbach & Furby, 1970; Stanley, 1967). For ease of understanding, we present the simplest formula (taken from Chiou & Spreng, 1996; also see Guilford, 1954 and Lord, 1963), which is as follows:

$$\rho_{dd'} = \frac{(\rho_{xx'} - \rho_{yy'})/\,2\, - \rho_{xy}}{1 - \rho_{xy}}$$

$\rho_{dd'}$ = estimated reliability of the difference score
$\rho_{xx'}$ = estimated reliability of component score A
$\rho_{xy'}$ = estimated reliability of component score B
ρ_{xy} = correlation between the two component scores

To futher aid the discussion, if we assume that the two component scores have equal reliability, the formula becomes:

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

$\rho_{dd'}$ = estimated reliability of the difference score
$\rho_{xx'}$ = estimated reliability of the component scores (equal reliability is assumed)
ρ_{xy} = correlation between the two component scores

To unpack this formula, let us first look at two extreme cases. The first is when each of the component scores is perfectly reliable (ρxx’ = 1). In this hypothetical case, the resulting difference score will be perfectly reliable because the numerator and denominator will both be the same. The second case is when the component scores are completely independent (ρxy = 0). In this case, the resulting reliability of the difference score will be equal to the reliability of the

component scores (which we assume to be equal here). Thus, if the component scores are perfectly reliable, then so too will the difference score. And if the component scores are completely independent, the difference score will be as reliable as the component measures. In both of these instances there is no reason to be concerned about the reliability of the difference score, at least any more than the reliability of the components. However, these conditions are not practical. Behavioral science measures never perfectly reliable or completely independent. Further, a difference score represents the difference between performance on two highly related processes – it would be unclear how one would even interpret a difference between two completely independent measures or why a researcher would want to create a difference score under that condition. So, ignoring the nonsensical cases of perfect reliability or complete independence of the component scores, **difference scores are necessarily less reliable than their individual components.**

The primary issue is that as the *correlation* between the two component scores *increases*, the *reliability* of the resulting difference score *decreases*. This is demonstrated in Figure 1 using hypothetical data and reasonable assumptions about the reliability of component scores (the three curves represent different levels of component score reliability). Importantly, when the correlation between the component scores approaches the extent to which their reliability permits, the resulting difference score is almost completely unreliable. For example: if mean reaction time on the incongruent trials and congruent trials of a Stroop task have a reliability of .90 and correlate with one another at r = .80, the resulting reaction time interference effect (difference score) will only have a .50 reliability. If the mean reaction time on the incongruent and congruent trials of a Stroop task have a reliability of .80 and correlate with one another at r = .70, the resulting difference score will have a reliability of .33. Now we can begin to see why the

Stroop effect, while being universal and robust at the group level, typically has low reliability and does not strongly correlate with other measures.

<Figure 1>

Conceptually, the lower reliability of difference scores is due to the correlation between component scores subsuming the bulk of their reliable variance. When two variables correlate, it means they share some amount of systematic variance. Because this variance is common to both variables, subtracting one of these variables from the other necessarily removes some of this systematic (reliable) variance. Subtraction therefore removes at least some reliable variance present in the two variables, and increases the proportion of error variance in the resulting difference score (e.g., Cronbach & Furby, 1970; Hedge et al., 2018). This explains why there is no issue with difference score reliability if component scores are perfectly reliable – the proportion of error variance cannot increase (relative to reliable variance) if no error variance exists to begin with. Similarly, if two variables are completely independent, there is no shared variance between the two component scores to subtract out, and the resulting difference score is equally reliable as the components. As a consequence to subtraction increasing the proportion of error variance, difference scores are prone to unreliability and typically have weak associations with other variables.

For the reasons outlined above, some researchers argue against the use of difference scores entirely. Lord (1956) observed, “differences between scores tend to be much more unreliable than the scores themselves,” (p. 429). In a subsequent paper, he notes “the difference between two fallible measures is frequently much more fallible than either,” (Lord, 1963, p. 32)

and that difference score procedures often produce absurd results. Cronbach and Furby (1970) emphatically stated:

> … scores formed by subtracting pretest scores from posttest scores lead to fallacious conclusions, primarily because such scores are systematically related to any random error of measurement. Although the unsuitability of such scores has long been discussed, they are still employed, even by some otherwise sophisticated investigators. (p. 68)

They also remarked, "It appears that investigators who ask questions regarding gain scores would ordinarily be better advised to frame their questions in other ways" (p. 81). Noting that difference scores are *sometimes* numerically reliable, Edwards (2001) adds that it is not a question of whether difference scores are reliable absolutely, but whether they are more reliable than other options. He also argues, "Moreover, adequate reliabilities do not absolve difference scores of their other methodological problems, and these problems are sufficient to proscribe the use of difference scores regardless of the reliabilities they exhibit" (p. 267). More recently and within the domain of executive functioning, Paap and Sawi (2016) argued, "… difference scores have low convergent validity that is partly caused by deficiencies in test-retest reliability" (p. 81).[7]

Arguments in favor of difference scores

As Collins (1996) attested, "there are few topics in social science methodology that have elicited as much confusion, misunderstanding, and anxiety as . . . gain scores" (as cited in

[7] There are additional considerations related to difference scores and the measurement of change in general that we do not discuss here. For instance, Sriram, Greenwald, and Nosek (2010) argue that factors which influence reaction time in general (subject characteristics, task demands, and the interaction between the two) cause correlations between reaction time difference scores to be biased and difficult to interpret. Campbell and Kenny (1999) discuss regression to the mean effects in measuring change.

Bezruczko, Fatani, & Magari, 2016, p. 289). Yet, researchers continually employ difference scores, and some contend that their methodological problems are exaggerated or misunderstood.

Tisak and Smith (1994a) judged difference scores to be an acceptable dependent variable **when the component scores are reliable and are not highly correlated with each other**. Zimmerman and Williams (1982) argued that unequal variance and reliability of the component scores is common, and that difference scores often have perfectly acceptable reliability when this is the case. However, they also conceded that, "It is undoubtedly true that many gain scores and difference scores are unreliable" (p 67). Chiou and Spreng (1996) similarly argued:

> While researchers may need to be cautious about the reliability issues of difference or gain score [sic] … in many practically possible situations, difference scores can still be very reliable. The purported unreliability of difference scores is partly due to unrealistic assumptions of the classical reliability formula. (p. 158)

Regarding Chiou and Spreng (1996) and Zimmerman and Williams (1982), it is important to again note that the formula we provide for the calculation of difference score reliability assumes equal reliability and variance of the component scores. This is assumed because a difference score is typically a variable on the same test measured either at two points in time (e.g., gain scores) or in two different, but similar, conditions (e.g., incongruent and congruent Stroop trials). The extent to which these assumptions do not hold can impact the reliability, as Chiou and Spreng (1996) and Zimmerman and Williams (1982) showed. Specifically, these researchers acknowledged the legitimacy of concerns regarding difference score use, but they emphasized that equal reliability and variance of component scores is not a given. When these assumptions do not hold, a formula with additional terms is needed to

calculate difference score reliability (Equation 1 in Chiou & Spreng, 1996), and difference scores can be shown to be somewhat more reliable. Rogosa and Willett (1983) also explored how the relationship between initial status and change affects the reliability and appropriateness of difference scores. Like Chiou and Spreng, they showed that unequal reliabilities between the component scores can lead to a more reliable difference score. Relatedly, Gollwitzer, Christ, and Lemmer (2014) argued that *in pre- and post-test treatment designs,* variance is more likely to be different across the two administrations (i.e., different standard deviations of the component scores) due to differential treatment effects.

Another complication that contributes to confusion over difference score use is the counterintuitive finding that unreliable scores can be beneficial for experimental researchers because the power of statistical tests is actually increased when the reliability of the difference score is low (and maximized when the reliability is zero; see Overall & Woodward, 1975). Chiou and Spreng (1996) explained one way this can happen:

> If there is no treatment by subject interaction effect (every subject shows the same true change), then based on the definition of the reliability formula, the difference score will be very "unreliable" even if the pretest and post-test score are reliably measured. This unreliability is caused by the low variation of the difference score across subjects. But not because of the measurement errors. Under this situation, the difference score is just like a constant. It is not suitable for correlation or LISREL analysis. This "unreliable" difference score, however, can be very powerful in a statistical test for a main effect (ANOVA or t-test) for within group design as long as the pretest and post-test score are reliably measured. (p. 164)

Finally, Trafimow (2015) argued that the observed correlation between the two components of a difference score is affected by their true correlation and that the interaction between component reliabilities with the true correlation largely determines the reliability of the resulting difference score. When variances and reliabilities of the component measures are unequal, the interaction between these variables gets more complex. He concluded:

> Does it matter whether the difference scores are reliable? That depends on one's purpose. If the goal is simply to demonstrate that the treatment works, lack of reliability of the difference scores likely will not be fatal for the statistical test, as Thomas and Zumbo (2012) showed. **But if the goal is to correlate the difference scores with another variable, then difference score reliability will matter a great deal** [boldface added]. (p. 10)

To sum, difference scores are contentious and statistically more complex than at first glance. Psychometricians have historically argued against their use from a mathematical perspective and given a certain set of assumptions (e.g., Cronbach & Furby, 1970; Lord, 1963). Others are not as categorically opposed to difference scores and argue that these assumptions are not always tenable, and that difference scores can be highly reliable in various situations (e.g., Tisak & Smith, 1994a; Zimmerman & Williams, 1982), which researchers must assess on a case-by-case basis (Trafimow, 2015). Just as with reliability, the critical distinction is the intended use of the difference score – experimental or differential. Defenders of difference scores are primarily concerned with their use in *experimental* designs. Their position is that the experimental researcher should not be too concerned about using difference scores **so long as experimental effects consistently emerge**. However, differential and developmental researchers

are not concerned with maximizing statistical power for ANOVA and t-tests, but rather in maximizing *reliable* between-subject variance. Difference scores are poorly suited for this purpose. Our position is that the all-too-common absence of a distinction made between experimental and differential approaches has largely contributed in the confusion and contention regarding difference score use.

Empirical assessment of difference score reliability in the context of individual differences in executive functioning

Given the conflicting views regarding difference scores, to what extent should differential researchers be concerned about their use? The research we discuss in the following sections suggests that, at least in the context of assessing individual differences in cognition, difference scores are better avoided.

Note that we limit this discussion to reaction time-based differences. Accuracy-based difference scores are often much less reliable than reaction time ones because they suffer the same methodological problems of reaction time difference scores but are also much more susceptible to restricted variance and ceiling effects. But, we do not discuss accuracy-based difference scores because, 1) reaction time is a problematic variable for other reasons than just in the form of a difference score (see the following sections), and, 2) accuracy-based difference scores are used much less frequently than reaction time differences, at least in our area of study. Many notable paradigms in psychology rely on reaction time difference scores, but we are aware of many fewer paradigms that both utilize *accuracy* differences and are prevalent in individual differences research.

While we only present examples from within cognitive psychology, we refer readers interested in how difference scores may affect other areas of research to the list we presented

previously on the ubiquity of difference scores. For examples of reviews and controversies of difference scores in other fields, see Gottman & Rushe (1993) for clinical psychology; Edwards (1996; 2001) for organizational behavioral research; Collins (1996) and Gollwitzer, et al. (2014) for social psychology; Peter, et al. (1993) for marketing research; Kessler (1977) for longitudinal survey research; DeGutis et al. (2013) for individual differences in holistic face processing; and Cafri et al. (2010) for psychiatric body image work.

Paap and Sawi (2016) recently evaluated the test-retest reliabilities of several popular executive functioning tasks that are predicated on difference scores (N = 81 university students). The pure reaction time (component score) reliabilities were generally quite high (.71 - .89), but the resulting cost scores and interference effects (difference scores) had test-retest reliabilities in the .43 - .62 range, meaning that roughly 50% of the variance in these difference scores were unreliable across the two administrations.

Salthouse, Fristoe, McGuthry, and Hambrick (1998) investigated the relationship between task switching, age, processing speed, and fluid intelligence. In their first study (N = 100 undergraduates), their three task switching measures had reaction time switch costs with reported Spearman-Brown corrected split-half reliability estimates of .71, .46, and .6 (Brown, 1910; Spearman, 1910). 1. While these were somewhat more reliable than Paap and Sawi (2016), this is to be expected because they are internal consistency estimates as opposed to test-retest reliability. The baseline reaction times (component scores) had very high reliabilities ranging from .91 - .95, which also likely contributed to the components being more reliable (recall from the difference score reliability formula that highly reliable component scores can lead to more reliable difference scores). Their second study had older adults perform the same switching tasks (N = 161 adults aged 18 – 80), and reaction time switch costs on them had reliabilities of .38,

.38, and .41. Baseline reaction time (component scores) had very low reliabilities of .64, .50, and .36, perhaps due to their inclusion of individuals in the 70 – 80 age range.

Siegrist (1997) assessed the reliability of Stroop performance by testing 45 undergraduate subjects on different variations of Stroop trials: "XXXXX" strings, conflicting color words, three types of self-relevant words, and two types of taboo words. The individual composites had test-retest reliability estimates in the range of .84 - .91, but the resulting interference effects (difference score) had test-retest reliabilities at .68, .09, -.12, and -.04. As a correction, Siegrist subtracted only "XXXXX" trials from color, taboo, and two types of self-relevant trials to produce interference effects with estimated test-retest reliabilities of .68, .48, .53, and .54, respectively. Siegrist concluded that these were all adequate because they reached statistical significance. However statistical significance is not sufficient to demonstrate acceptable reliability (see Bonett, 2002; Nunnally & Bernstein, 1994; Schönbrodt & Perugini, 2013), just as statistical significance does not demonstrate the meaningfulness of an effect. Furthermore, his estimates are likely inflated as he compared trial types that were not wholly similar to one another (XXXXX trials from other types of trials), resulting in weaker composite score correlations and slightly more reliable, but less interpretable, interference effects. As such, Siegrist's reliability estimates were in the upper range of what one would expect with Stroop interference effects. Even with this inflation, they still falls well short of Nunnally's (1964) .80 recommendation for reliability in basic research.

Whitehead, Brewer, and Blais (2018) assessed the reliability of congruency effects in a color Stroop, Simon, and letter flanker task (just under 200 subjects in each of three experiments). The sequential congruency effect is the finding that interference effects (e.g., slowing and reduced accuracy on incongruent trials of the Stroop as compared to congruent

trials) are smaller when preceded by another trial of high interference (Gratton et al., 1992). As such, an incongruent trial preceded by a congruent trial will result in a larger interference effect than an incongruent trial preceded by another incongruent trial. Sequential congruency effects are assessed with a difference between two interference effects – **a difference of two difference scores**. Whitehead et al. (2018) report that sequential congruency effects in each task had effectively no reliability in any of the three experiments. Even-odd split half estimates with Spearman-Brown correction ranged from -.07 to .17, with only one of the nine being statistically significant. Because reliability constrains validity, performance on these measures subsequently showed no statistically significant inter-correlations. Yet, these sequential congruency effects were present at the group level. These results are similar to those Logie et al. (1996) discussed previously and in line with Overall and Woodward's (1975) demonstration that statistical power of ANOVA tests is increased in measures of low reliability. Whitehead et al. also measured post-error slowing in these tasks and reported higher reliabilities for post-error slowing (.47 - .84, with most measures being in the .60s), with post-error slowing correlating across the three tasks (r = .29 - .50). This finding highlights that difference scores *of* difference scores cannot be expected to produce any reliable individual differences since the issue with difference scores is compounded. It also reinforces that robust experimental (group-level) effects do not necessarily produce reliable individual differences. This series of experiments also illustrates another important nuance of the difference score issue; depending on one's research question, they may occasionally be useful even in individual differences studies. While only around 65% of the variance in post-error slowing was reliable within each task, this was reliable enough for Whitehead et al. to find that post-error slowing correlated across tasks and thus was not task-specific. However, if Whitehead et al. were concerned instead with the *magnitude* of the true

correlation of post-error slowing across tasks and not just the presence of this association, then using these difference scores with reliability in the .60s would be a severe limitation.

Our research team typically uses the color Stroop and arrow flanker tasks as attention control measures. In two recent large-scale studies, reaction time on congruent and incongruent trials (54 trials for each type) correlated at r = .86 - .89 in the Stroop and .87 - .88 in the arrow flanker, resulting in difference scores with Spearman-Brown corrected split-half coefficients below .70 in all cases. Data from two task switching procedures showed a similar pattern of high correlation between component scores (here switch and repeat trials), which resulted in marginally reliable difference scores at .64 and .73. Comparing the difference score internal consistencies to our accuracy-based executive functioning measures is striking, as difference score measures had demonstrably lower internal consistency in both studies (see the Appendix).

The data outlined above show how component scores (e.g., incongruent and congruent trials in the Stroop) are typically highly correlated and produce difference scores with much lower reliability than the components themselves. Component measures for the tasks used to assess executive attention and executive functioning consistently correlate around r = .90, with the resulting difference scores often having test-retest reliability and internal consistency estimates below .70. Therefore, Tisak and Smith's (1994a) argument that difference scores are not problematic given weakly correlated components does not apply.[8]

[8] Wöstmann et al. (2013) reported internal consistencies and test-retest reliabilities of numerous executive functioning measures such as the Eriksen flanker, Simon task, and Stroop task. While *some* of their reliability estimates were quite large (.53 - .94 for the flanker interference effect and .69 - .89 for the Simon effect), we excluded their study from discussion because it had a sample size of only 23, falling well below the recommended minimum for assessing reliability (see Bonett, 2002; Nunnally & Bernstein, 1994; Schönbrodt & Perugini, 2013). We considered removing any discussion of Siegrist (1997) for similar reasons, however the Siegrist data are informative because they show how comparing incongruent and congruent trials from different tasks can produce more reliable differences, but at the cost of interpretability.

Importantly, the results discussed in this section are not anomalous cases. Rather, these are illustrative examples of the psychometric issues with many paradigms relying on reaction time differences. And these are issues that other researchers have acknowledged, both within the measurement of attention and task switching (e.g., Friedman & Miyake, 2004; Hughes et al., 2014; Rey-Mermet et al. 2018; Vandierendonck, 2017; 2018) as well as other areas and disciplines (e.g., DeGutis et al., 2013; Collins 1996; Edwards, 1993; Gottman & Rushe, 1993; Peter, et al., 1993). Our position is that these measurement issues stem from the use of reaction time scores, especially reaction time differences, and that the problems with these scores in individual differences contexts are pervasive throughout psychology and behavioral research more broadly. In following sections, we provide a more in depth discussion of these issues as they pertain to specific areas of cognitive psychology.

Reaction time measures are sensitive to speed-accuracy relationships

So far, the issues we have discussed regarding reaction time have been that reaction time differences scores do not result in reliable individual differences. However, reaction time *in general* is a concern because it is sensitive to speed-accuracy interactions that may differ across ability and developmental levels.

Speed and accuracy can interact in a number of ways. Most important for present purposes, emphasizing one response dimension (speed or accuracy) often results in a detriment for the other, known as the *speed-accuracy tradeoff* (see Heitz, 2014 for a review). Speed-accuracy tradeoffs are often considered a nuisance because researchers are typically interested in either error rates *or* reaction time and assume the other is unimportant, but the two are intimately connected and not always in a predictable or obvious manner.

When a researcher employs a reaction time measure, they assume that *all* differences in performance manifest through reaction time. For this assumption to be valid, all respondents would have to emphasize speed and accuracy to the same degree. To ensure this, subjects are customarily instructed to, "Respond as quickly and accurately as possible," which, given the inverse speed-accuracy relationship, is ambiguous and contradictory (see Edwards, 2001 and Heitz, 2014). While instructions can be made clearer, individuals will adopt different response criteria regardless of speed-accuracy instructions (e.g., Heitz, 2014; Lohman, 1989), and that subjects are prone to initially modifying their performance only to later revert to their more natural response tendencies (e.g., Schouten and Bekker, 1967). Individuals will continue to interpret and respond to these instructions differently, and so instructions alone are not sufficient to equate individuals in terms of speed-accuracy tendencies. More to the point we wish to make, subjects of different ability and developmental stages will likely respond differently to these instructions.

Speed-accuracy interactions are a problem in differential and developmental studies because minor variations in emphasis on speed versus accuracy across subjects threatens the reliability and validity of a measure. Reaction time measures are especially susceptible because accuracy is often ignored altogether in reaction time analyses. Some individuals will likely respond impulsively or slowly regardless of instructions (see Starns & Ratcliff, 2010). Furthermore, individuals of differing ability levels likely differentially adjust speed and accuracy to meet task demands and instructions. Higher ability individuals may slow down if they notice they are making errors, whereas lower ability individuals may not (e.g., Draheim, Hicks, & Engle, 2016). Complicating matters further is that individual differences in speed-accuracy adjustments can, but do not always, emerge. For instance, Unsworth, Redick, Spillers, and

Brewer (2012) report that microadjustments (including post-error slowing) were present across four different cognitive control tasks, but that these microadjustments were not different across individuals of differing working memory capacity (perhaps due to the limitations of difference scores in individual differences contexts).

Additional concerns related to the speed-accuracy relationship is that the tradeoff between the two is asymmetrical in that sizeable changes in reaction time will likely produce very minor, perhaps even undetectable, changes in accuracy rates (Forstmann et al., 2011; Pew, 1969). Forstmann et al. (2011) claim that, due to their asymmetrical relationship, "… traditional methods of inference … may overlook the effects on accuracy and mistakenly conclude that speed-accuracy tradeoff differences do not play a role" (p. 17242). The complex interplay between speed and accuracy can thus lead to misleading conclusions in analyses based solely on reaction time (see Draheim et al., 2016; Regev & Meiran, 2014), and so speed-accuracy interactions are both highly prevalent and highly problematic in differential research involving reaction time. This is particularly true in studies with diverse and heterogeneous samples, as is common in individual differences studies, developmental studies, aging studies, and clinical research, in which there is a higher likelihood of finding individual or group differences in speed-accuracy relationships. For example, it has been repeatedly shown that older adults consistently favor slow and accurate responding by default and are also typically either unwilling or unable to sacrifice accuracy for the sake of speed regardless of practice, instructions, or incentives (e.g., Brébion. 2001; Botwinick & Storandt, 1973; Forstmann et al. 2011; Hertzog, Vernon, & Rypma, 1993; Rabbitt, 1979; Salthouse, 1979; Salthouse, 1996; Starns & Ratcliff, 2010). Aging research has also revealed that trial-to-trial and day-to-day reaction time variability is important in understanding cognitive processing, and that individuals of differing abilities

often show different variability in reaction time (e.g.,Hertzog, Dixon, & Hultsch, 1992; Hultsch, MacDonald, & Dixon, 2002; Rabbitt, Osman, Moore, & Stollery, 2001). In particular, the slowest 10 or 20% of trials is often more informative and indicative of ability than the rest of the reaction time distribution, notably in sustained attention paradigms. For this reason, some reaction time tasks are scored as the average of an individual's slowest *N* percentage of trials instead of their mean reaction time across all trials (Dinges & Powell, 1985; Unsworth & Robison, 2016). Many aging researchers thus have a heightened awareness of the potential issues with assessing performance with reaction time, and how factors such as speed-accuracy interactions and intraindividual variability need to be addressed. Unfortunately, researchers in other areas of psychology tend to be less cognizant and only consider these intricacies when reaction time data yield unanticipated or null results.

To summarize, speed-accuracy interactions are important to consider in differential and developmental research because such studies, by design, involve the testing of subjects with a wider array of ability levels – making the presence of differences in speed-accuracy tradeoffs and related strategies inevitable. And, just like the researchers who argue against using reaction time difference scores, some researchers argue that simple reaction time measures are inappropriate due to their susceptibility to speed-accuracy interactions (e.g., Forstmann et al., 2011; Luce, 1986; Wickelgren, 1977). These researchers believe that speed-accuracy tradeoffs require consideration in most reaction time research, and they championed techniques to more directly study speed-accuracy tradeoffs. For example, Luce (1986) stated:

> … we face a very common problem in psychology: the existence of a tradeoff between dependent variables, in this case false alarms and reaction time. The only

> sensible long-term strategy is, in my opinion, to study the tradeoff… and to devise some summary statistic to describe it. (p. 56-57)

There is a large body of work devoted to doing just this, with the systematic study of speed-accuracy tradeoffs exploding in the mid-to-late 1960s (e.g., Fitts, 1966; Ollman, 1966; Pachella & Fisher, 1972; Pachella & Pew, 1968; Schouten & Bekker, 1967) coinciding with the idea that a single summary statistic (e.g., reaction time) is insufficient to measure cognitive processes. Wickelgren (1977) in particular argued that methods of either separating or manipulating speed-accuracy interactions are so superior to traditional reaction time measures that cognitive psychologists should consider them the default. We will briefly review some of these speed-accuracy methods in our section dedicated to alternatives to reaction time, but we direct interested readers to Heitz (2014) for a more extensive review.

The impurity of reaction time correlations

Another complexity to reaction time comes from Miller and Ulrich (2013). They argued that the cognitive processes underlying reaction times are more complicated than researchers typically assume, leading to faulty interpretations of reaction time scores. They proposed a model for predicting individual differences in mean reaction times. The model consists of a series of task-specific information processing steps assumed constant across individuals insofar as the same task requires different individuals to execute the same amount of mental work. Between-subjects variability is introduced via differences in individuals' abilities to efficiently complete said work. The model is displayed below:

$$RT_k = (A + B + C) \times G_k + B \times \Delta_k + R_k + E_k$$

Briefly, terms A, B, and C correspond to the amount of time needed for different information processing stages required by a task. Terms A and C correspond to perceptual input and motor output stages, while term B represents task-central processes such as decision making, response selection, and information manipulation. The model's other terms represent idiosyncratic processing times, including general processing speed (G_k), processing speed on the central task, i.e., stage B (Δ_k), residual differences in speed unrelated to G_k or Δ_k such as those associated with stages A and C (R_k), and differences due to random error (E_k).

The model and its variations have far-reaching implications for interpreting reaction time reliabilities and correlations. Miller and Ulrich (2013) reported that reliabilities for a simple mean reaction time should be satisfactory given sufficient sample variation in any one of the G, Δ, or R terms. However, their model suggests that strong reliability estimates are of limited importance for interpreting reaction time scores, since it is an open question as to which parameter(s) exert influence on the reliability estimates, and to what extent. For example, variation in the residual term R would yield reliable reaction time scores. However, most research aims at understanding the central processing stage, B, and would be most interested in variation in the Δ term. A reaction time measure that is only reliable due to variation in the R term would be of little interest to most researchers, since it would be unlikely to correlate with other measures in theoretically informative ways. In short, the process impurity of reaction time measures is a major hindrance to their empirical and theoretical utility. Paap & Sawi (2016) take a similar stance on this issue and note the conundrum this creates for researchers: difference scores are preferable to simple reaction times because they partially mitigate the impurity issue, but difference scores are notoriously unreliable and not suited to individual differences studies.

Importantly, Miller and Ulrich's (2013) model does *not* suggest that correlations derived from reaction time measures will necessarily be weak, but rather that these correlations are unlikely to be informative. Their model suggests that, if anything, simple reaction time correlations are prone to inflation given that general processing time (G) will be similar across a wide array of tasks. This becomes more complicated when working with difference scores, in which the problems with process impurity are compounded, and predicted reliabilities and correlations can vary widely depending on the relationship between a difference score's components and their idiosyncratic parameters. Miller and Ulrich (2013) thus concluded that "…there is a long way to go before it will be possible to draw strong conclusions from the size or in some cases even the direction of a reaction time-based correlation" (p. 839), and that, "… relatively sophisticated research strategies will be required to reach strong conclusions from between-task correlations of mean reaction times" (p. 840).

Validity Concerns of Reaction Time Measures

Reaction time is so prevalent in behavioral research that it not possible to discuss all the fields and paradigms affected by the issues that we have discussed thus far. However, we have selected a few areas within cognitive psychology in which reaction time differences have led to controversial findings. Our position is that null and inconsistent findings are primarily due to a combination of the unreliability of reaction time difference scores, reaction time being sensitive to speed-accuracy interactions, and, in some cases, the aforementioned issues raised by Miller and Ulrich (2013) regarding the impurity of reaction time.

Task switching

The relationship between task switching and working memory has been a contentious topic recently and is a good example of how difference scores can be useful for experimental

researchers but problematic for differential purposes. Task switching is operationalized as the ability to flexibly and fluidly switch attention and other cognitive resources from one task to another to meet task demands (see Jersild; 1927; Monsell, 2003), and interest in it proliferated in the mid-1990s (e.g., Allport, Styles, & Hsieh, 1994; Meiran, 1996; Rogers & Monsell, 1995). Modern task switching paradigms are used to study cognitive control and executive functioning (e.g., Altmann & Gray, 2008; Miyake et al., 2000; Oberauer, Süß, Wilhelm, & Wittman, 2003), and, as Logan (2004) proclaimed, "The ability to switch flexibly between tasks is the pinnacle of human cognition and the hallmark of executive control" (p. 220).

A consistent finding from the task switching literature is that individuals exhibit slower and more error-prone responses on trials in which they must change their rules for responding (switch trials) than for trials in which a switch is not required (repeat or non-switch trials).[9] A well-accepted theoretical account is that a memory representation of the configuration of rules for performing each task (the task set) must be maintained in a readily accessible form to be retrieved when a switch is required. Rogers and Monsell (1995) argue that switch costs arise due to a task set reconfiguration process on switch trials, whereas Allport et al. (1994) argue that switch costs reflect proactive interference from previously active and relevant but now irrelevant task sets (for reviews, see Monsell, 2003; Kiesel et al., 2010; Vandierendonck, Liefooghe, & Verbruggen, 2010). There is debate as to the extent to which working memory is involved in task switching. Some researchers believe that working memory completely mediates task switching performance (e.g., Mayr & Kliegl, 2000; Rubinstein, Meyer, & Evans, 2001), but others

[9] Task switching tasks usually involve a series of simple categorization or judgment task with two potential options such as even/odd, large/small, living/non-living. A switch trial is one that requires the subject to make a different judgment than previously, for example if an object is living/non-living whereas the previous trial required a large/small judgment. A repeat trial is one that requires the subject to make the same judgment as on the previous trial.

implicate processes outside of working memory, such as long-term memory (e.g., Allport et al., 1994; Logan & Gordon, 2001). However, both camps agree that working memory is a vital process in order to switch between tasks (e.g., Mayr & Keele, 2000), and it has been repeatedly demonstrated *experimentally* that switching between tasks taxes working memory (e.g., Baddeley, Chincotta, & Adam, 2001; Emerson & Miyake, 2003; Liefooghe, Barrouillet, Vandierendonck, & Camos, 2008; Liefooghe, Vandierendonck, Muyllaert, Verbruggen, & Vanneste, 2005).

Despite the theoretical acceptance and experimental support for the relationship between task switching and working memory capacity, several factor analytic studies have failed to find this link. Notable studies include Miyake et al. (2000), Oberauer et al. (2003), and data from our research team discussed in Draheim et al. (2016). Oberauer et al. found little relationship between task switching and working memory capacity and concluded:

> Supervision, as operationalized by the task set switching variables, was only weakly related to the other working memory functions. To the extent that switching reflects a function of the central executive in terms of Baddeley's (1986) model, this implies that at least some aspects of the central executive are not very central to working memory. Our result is in accordance with Miyake et al. (2000) who found little relationship between their mental set shifting factor and standard measures of working memory capacity… (p. 189-190)

Our results were even more puzzling. We found a moderate relationship between working memory capacity and task switching, but in the opposite direction predicted by theory - individuals with higher working memory (and fluid intelligence) scores had *larger* switch costs,

suggesting that higher ability individuals were worse than lower ability individuals at switching between tasks.

The failure to demonstrate a meaningful correlation between working memory capacity and task switching suggests either that the theory of at least one or both of the two abilities is flawed, or that the measurement of at least one of the abilities is at issue. Miyake et al. and Oberauer et al. argued that their results had significant theoretical implications, but we attributed the results to the measurement of task switching. Researchers involved in these three studies employed a wide array of task switching tasks, assessed working memory capacity (and other executive functions) differently, and tested very different populations. Yet we all failed to find the theoretically predicted strong and positive relationship between task switching and working memory capacity. Crucially, however, we all assessed task switching using reaction time difference scores. Given this commonality, the problems we discussed previously with difference scores, and that experimental studies have demonstrated a link between working memory and task switching, it is likely that the measurement of task switching was the issue.[10]

We found evidence for this position in a reanalysis of the data from our data and Oberauer et al.'s (2003) data using a measure that incorporates both speed *and* accuracy, and with less dependence on differences in reaction time (Hughes et al., 2014). This reanalysis of Oberauer et al.'s (2003) data revealed a much stronger relationship between task switching and working memory capacity than was initially apparent, and this relationship manifested at the individual task, the composite, and the latent levels. When we applied the integrative speed-accuracy measure to our own data, the weak-to-moderate negative relationship between the two constructs we initially found when we assessed task switching with reaction time switch costs

[10] To reiterate, our position is that the measurement of task switching is problematic for differential research, but studies have consistently shown switch costs to be very useful in experimental contexts.

became very strong and positive, on the order of around r = .50 at the composite level. The reanalysis thus supported prior theoretical and experimental work on the relationship between task switching and working memory capacity, whereas reaction time difference scores (or reaction time alone) revealed either null results or results that contradicted theory.

Task switching is a prime example of how difference scores can lead to confusion in correlational research, both because of the unreliability of difference scores and differences in speed-accuracy interactions across the cognitive ability spectrum. In Draheim et al. (2016), individuals with higher working memory scores tended to slow down after errors, as there was a significant correlation between working memory capacity (and fluid intelligence) and the extent to which an individual slowed down on a trial immediately following an error. This finding illustrates the danger in emphasizing results that are plagued with both speed-accuracy interactions and unreliable scores. The interpretation of the data changed drastically when we accounted for speed-accuracy interations, in this case with an integrative measure of speed and accuracy, *and* when we did not rely on difference scores. The result was more sensible and in line with established theoretical accounts of the constructs in question and congruent with the experimental literature. In the alternatives section, we provide a more detailed discussion of integrative speed and accuracy measures.

Inhibition / Attention Control

Inhibition is an important component of executive functioning, though researchers disagree substantially over its nature and measurement (Logan, 1985). We regard true inhibition (i.e., the dampening or suppression of irrelevant information to the benefit of relevant information) to only exist in a few specific cases, such as lateral inhibition in the retina (e.g., Cook & McReynolds, 1998), but we acknowledge many attention control tasks do require some

inhibition-like processes. For example, in the complex span tasks, resisting proactive interference is a large determinant of performance, especially in later trials (Engle, 2002). In fluid intelligence tasks, disengagement from previously tested hypotheses results in better performance (Shipstead, Harrison, & Engle, 2016). And, in the antisaccade, subjects must resist the evolutionarily ingrained prepotent response to look toward the flickering distractor since flicker resembles evolutionarily-important movement. Things that move may be able to eat you, or you may be able to eat them. The color Stroop and arrow flanker tasks are also largely considered to measure inhibition. In the Stroop, subjects must resist the automaticity of reading and respond instead to the color of ink in which a word is presented. In a flanker task, respondents must focus on a single central target among multiple distractors (e.g., a central arrow pointing in the opposite direction of two flanking arrows on either side; Friedman & Miyake, 2004; Miyake et al., 2000; Rey-Mermet et al., 2018). It is likely that inhibition-like processes are necessary to some degree in performing most executive functioning tasks. As this article is not a review on the theory of inhibition and attention control, for the present purposes we consider *inhibition* and *attention control* to refer to the same concept.

Issues with assessing individual differences using attention tasks such as the Stroop and flanker are well documented. We report data for the reliability and correlations among our working memory capacity, fluid intelligence, and attention control measures in Table 2 and 3 of the Appendix. When we assess working memory capacity, fluid intelligence, and attention control, performance on most tasks correlate with each other strongly (typically r = .45 - .60) except with the Stroop and flanker reaction time interference effects, which correlate with other measures at r = .30 at the strongest and, on average, around r = .20. Accuracy rates on the antisaccade task, a hallmark measure of attention control (e.g., Hutton & Ettinger, 2006),

consistently correlate to working memory capacity and fluid intelligence measures more strongly than reaction time differences in the Stroop and flanker, despite consensus that the Stroop and flanker tasks are also attention control/inhibition measures (e.g., Heitz & Engle, 2007; Kane & Engle, 2003; Lavie, 2005; Miyake et al., 2000; Rey-Mermet et al., 2018; Unsworth & Spillers, 2010). A similar pattern emerges with performance on the visual arrays, a change detection task with both attention control and capacity requirements and an accuracy-based dependent variable (e.g., Luck & Vogel, 1997; Pashler, 1988; Redick et al., 2016; Shipstead & Engle, 2013; Shipstead, Lindsey, Marshall, & Engle, 2014). Specifically, visual arrays performance tends to correlate above $r = .40$ with all other attention and working memory measures except the Stroop and flanker reaction time interference effects (around $r = .20$). One could argue that this pattern of results is due to the Stroop and flanker being the only tasks with a reaction time dependent variable. However, the weakest correlation among all our variables is typically between the Stroop and flanker effects at around $r = .10$; **around only 3% of the total variance in performance is shared between them at the task level.** This is highly problematic given that these tasks are believed to measure related aspects of attention control (e.g., Friedman & Miyake, 2004) and have dependent variables calculated in a similar manner, and so a *much* stronger association between the two would be expected. On the other hand, high- and low-ability subjects (as indexed by working memory capacity) do show marked differences in error rates on these tasks, particularly on incongruent trials and in conditions with large congruent/incongruent trial ratios (see Engle & Kane, 2004 and Kane & Engle, 2003). Given the storied history of the Stroop being a classic experimental task that produces reliable and robust experimental effects (see MacLeod, 1991), this finding strongly indicates that the problem lies with the use of the

Stroop and flanker reaction time interference effects in correlational contexts, and not the tasks themselves.

Researchers are increasingly acknowledging the problems associated with the measurement of attention control (e.g., Friedman & Miyake, 2004; Hedge et al., 2018; Paap & Sawi, 2016; Rey-Mermet et al., 2018; Rouder & Haaf, 2018). Friedman and Miyake (2004) divided inhibition into three separate processes; inhibiting a prepotent response (e.g., antisaccade and Stroop), resisting interference from a distractor (e.g., flanker), and resisting proactive interference (e.g., cued recall and the Brown-Peterson task, see Kane & Engle, 2000) and tested 220 psychology undergraduates on tasks falling within these categories. Most of their measures were wholly unreliable. The resistance to proactive interference measures were so unreliable that they excluded them from their final analysis. First-order correlations among all measures were quite low, with more non-significant correlations than significant. Friedman and Miyake (2004) noted that difference scores likely factored into the low reliability and weak correlations of the measures. They also correctly stated that latent variable approaches can help correct for (but not absolve) this unreliability because only reliable variance is partitioned, though such methods are often impractical or even impossible because of their intense time and resource requirements.[11] Noting the psychometric issues of inhibition tasks, Friedman and Miyake (2004) concluded:

> One obvious solution to this problem is to develop new tasks that are psychometrically reliable and more sensitive to individual variation in inhibition-related processes. Although our strategy in the current study was to focus on existing measures used in the

[11] Even at the latent level, the relationship among the different processes of inhibition were suspect. This helps demonstrate that using factor analysis and structural equation modeling does not excuse the use of unreliable measures, as results can be unpredictable and untrustworthy, leading to erroneous conclusions.

> field, **it is becoming increasingly clear that new measures are needed for the field to make further progress…** (p. 127, boldface added)

Rey-Mermet et al. (2018) wrote that inhibition is an important topic in the aging literature because hypothesized age-related deficits in inhibition appear inconsistently (see Verhaeghen, 2011). Shadowing Friedman and Miyake (2004), they tested 130 young adults and 159 older adults on a battery of 11 inhibition tasks including two Stroop tasks, two flanker tasks, the Simon task, the stop-signal task, and an adaptive antisaccade task. All of their inhibition tasks were reaction time-based except for the antisaccade, and all of the reaction time-based measures were difference scores except for the stop-signal. The adjusted split-half internal consistency of their reaction time inhibition measures ranged between .27 - .85, with all but two (including the color Stroop) falling below .75 (see their Table 4), and yet these low reliabilities were higher than those found in many other studies. Their adaptive antisaccade task, on the other hand, had an internal consistency of .97. Despite using tasks which measure the same underlying ability, their 11 measures resulted in only 13 of 45 (29%) first-order correlations reaching statistical significance at the .05 level. Finally, they found low factor loadings for these tasks at the latent level.[12] These results led Rey-Mermet et al. (2018) to conclude:

> So far, the evidence suggests that the tasks used to assess inhibition do not measure a common underlying construct, but the highly task-specific ability to resolve the interference arising in that task … an inevitable implication of this conclusion is that studies using a single laboratory paradigm for assessing or investigating inhibition do not warrant generalization beyond the specific paradigm studied. (p. 515)

[12] Note that their antisaccade task also had weak correlations and poor factor loadings. We believe this to be because their antisaccade task was adaptive such that presentation rates were different for each subject depending on their performance during practice, resulting in less variability in error rate.

We agree with the sentiment that single tasks cannot properly measure a construct, and that doing so limits generalizability. We do not agree with Rey-Mermet et al. (2018) that their results warrant sweeping substantive claims about the unity of inhibition (attention control). Most of their inhibition tasks had low reliability stemming from a reliance on reaction time differences, which is a common issue in the measurement of inhibition. Their titular recommendation for researchers to stop thinking about inhibition as a general construct fails to take these methodological issues into consideration. It also ignores Friedman and Miyake's (2004) advice to create new, presumably non-difference score-based tasks before making such strong theoretical statements. As such, we express caution in interpreting their results – as our position is that their study is another example of psychometrically problematic measures misinforming psychological theory.[13]

Despite the consistent finding that performance on many reaction time-based attention tasks are unreliable and do not correlate with each other or external variables, many of these, especially Stroop and flanker tasks, see continued use in individual differences studies. A large reason we have continued to use them is that the reaction time interference effects (difference scores) on the color Stroop and arrow flanker cohere with error rates on the antisaccade to form an attention control factor with strong loadings to other cognitive constructs. However, the factor loadings for the Stroop and flanker tend to be very small whereas the antisaccade loading tends to be quite large (e.g., Kane et al., 2016; Miyake et al., 2000; Shipstead, Harrison, & Engle, 2015). We often observe Stroop and flanker performance (measured with reaction time difference scores) to have factor loadings in the .20s and accuracy on the antisaccade to have

[13] To be clear, we adjudge Rey-Mermet et al.'s conclusions to be correct *insofar as they are limited to the tasks they employed.* Our stance, however, is that better measures of inhibition would likely lead to much different, and more optimistic, results.

loadings in the .70s, indicating that our attention control factor is primarily composed of variance from only the antisaccade task. Rey-Mermet et al. (2018) similarly noted that inhibition factors are often dominated by a single measure. In their appendix, they listed 23 experiments across multiple labs that assessed attention control at the latent level. They determined 14 of these 23 experiments had an attention task that "dominated" the factor (high factor loadings for that one task and low loadings for the others). In nine of these 14 cases, that dominant task was a type of antisaccade task, and in only one experiment did another task dominate the factor when an antisaccade task was also part of that factor. In contrast, they listed two studies by Chuderski (2014; 2015) that exclusively used antisaccade tasks and found loadings at or above .80 for each of these tasks.

With the noted issues regarding the measurement of attention control and the idea that inhibition (attention control) may not be a unitary concept gaining traction (e.g., Rey-Mermet et al., 2018; Rouder & Haaf, 2018), it is important to keep in mind Friedman and Miyake's (2004) observation that new and improved tasks are needed in order to make any significant theoretical advancements. Our research team recently finished data collection for a study in which we tested over 400 individuals on a wide array of established, modified, and new attention tasks (Draheim, Martin, Tsukahara, Mashburn, & Engle, 2018). Preliminary results heavily suggest that threshold versions of the Stroop and flanker and accuracy-based tasks considerably improve the measurement of attention control over reaction time-based ones (color Stroop, arrow flanker, psychomotor vigilance task). Further, these data provide evidence that attention control is indeed a unified concept – so long as it is not measured with reaction time and reaction time difference scores.

Bilingualism

The role of lifelong bilingualism as the basis of enhanced executive functioning is another area of contention. A substantial body of work shows differences in performance in lifelong bilinguals compared to their monolingual peers, as well as neuroprotective benefits in older bilingual adults (Bialystok, 2017; Bialystok, Craik, & Freedman, 2007; Bialystok & Viswanathan, 2009; Luk, Bialystok, Craik, & Grady, 2011; See Li, Legault, and Litcofsky, 2014 for a review). However, some researchers consistently fail to show these effects using reaction time measures, and consequently argue that there is no executive function advantage from being bilingual (Paap and Greenberg, 2013; Paap, Johnson, & Sawi, 2015).[14] Moreover, studying bilingualism is further complicated by population differences in age of acquisition, manner of acquisition, and a reliance on self-report regarding bilingual status. Finally, the majority of these findings regarding the presence or absence of a cognitive change related to bilingual status used measures which rely on reaction time differences, likely impacting the replicability of results (e.g., Paap & Sawi, 2014).

Weinreich (1953) introduced the idea that interference between multiple languages occurs in bilingual individuals. He suggested that the presence or activation of multiple languages results in the need for bilingual individuals to resolve competition between them. Weinreich also suggested that this effortful selection could transfer generally to other cognitive processes. These suggestions were later supported by work by Costa, Caramazza, & Sebastain-

[14] Although there is heated debate about bilingual advantages in behavioral studies, neuroimaging methodologies more consistently report structural and functional differences between the brains of bilingual and monolingual individuals (Abutalebi & Green, 2016; Bialystok, 2017). However, we remain agnostic to the argument over the existence and nature of bilingual advantages and instead emphasize measurement problems which may contribute to the discrepant findings.

Galles (2000) and Costa, Santesteban, & Ivanova (2006), who showed language interference in very basic lexical tasks (see Kroll, Dussias, Bice, & Perrotti, 2015).

Green (1998) proposed a theory for how this joint activation and subsequent selection influences cognition. According to Green's inhibitory control model, a supervisory attention system guided by top-down cues inhibits the non-target language. Researchers reasoned that long-term use would strengthen these inhibitory processes, thereby enhancing inhibitory control in other non-linguistic domains. This inhibitory account has been investigated extensively, and became the primary explanation of the impact of bilingualism on cognition (Bialystok & Viswanathan, 2009; for a review see Kroll, Gullifer, McClain, Rossi, & Martin, 2015). Besides providing a plausible explanation for how bilinguals avoid interference between their multiple languages, the inhibition account was appealing because it accorded with contemporaneous advances in executive functioning theory (Miyake et al., 2000).

Many studies focused on bilingual advantages at this time have emphasized enhanced inhibition and/or switching abilities in bilingual individuals (e.g. Bialystok et al., 2009). Specifically, building on Green's (1998) inhibitory control theory, researchers frequently proposed that both languages are held active and that the unselected language was actively suppressed. This proposition has been extended to executive functions more broadly. and is supported by bilingual individuals showing advantages in task switching (Costa et al., 2000). However, suppression of the non-target language is not the only plausible explanation. Bialystok (1992; 2017) suggests that bilingual individuals may not be suppressing the non-target but are instead actively selecting the target language. This view moves the study of bilingualism away from an executive function approach toward an executive attention perspective more similar to that of Engle (2002; 2018). While the executive function approach takes a more deconstructive

approach in delineating specific, isolable, functions of the central executive in the working memory system, the executive attention perspective takes a broader approach to individual differences in performance rooted in the domain-general ability to control attention (Engle, 2002; 2018).

The debate in bilingualism research is intimately connected with our previous discussion of the controversy surrounding task switching measurement (e.g., Hughes et al., 2014) and inhibition (Rey-Mermet et al., 2018). The executive functioning approach has typically relied heavily on reaction time difference score measures whereas the executive attention approach has adopted more reliable accuracy based measures of attention control. The reliance on unequally reliable measures between these two positions makes comparison essentially impossible. We argue that reaction time difference scores along with the interaction of speed-accuracy emphasis with ability and developmental level may be partly to blame for the conflicting conclusions regarding the existence and nature of enhanced cognitive functioning in bilinguals. Further, even if, as some recent meta-analyses suggest (Donnelly, Brooks, & Homer, 2015; Lehtonen et al., 2018) the bilingual advantage is not a true effect, adopting more psychometrically rigorous measures will hasten the field toward sounder conclusions about this globally important question.

Sequential Learning

The serial reaction time task is a simple respond-to-cue task that has blocks with cues appearing randomly and other blocks with cues appearing in a set sequence. Differences in reaction time on the random trials vs. sequence trials are widely used by cognitive and neuroscience researchers to measure sequential learning. These differences are believed to indicate whether the learning is implicit or explicit. Because these reaction time difference scores

rarely correlate to measures of higher cognition, serial reaction time tasks are inferred to measure implicit rather than explicit learning (Urry, et al., 2015), as explicit learning would be expected to correlate with cognitive ability (e.g., Unsworth & Engle, 2005). Some researchers have naturally raised methodological concerns regarding this task due to the reliance on reaction time difference scores (e.g., Howard & Howard, 1992; Kaufman et al., 2010).

Urry et al. (2015) argued that difference scores on the serial reaction time task are unreliable, susceptible to floor effects, and are a theoretically inappropriate way to measure learning - particularly if accuracy is not taken into consideration. They developed a new sequential learning task designed to produce both reaction time and accuracy dependent variables. They assessed performance on this task as well as a traditional serial reaction time task (N = 99, undergraduates and community members) using a reaction time difference score, a ratio reaction time score (which reflects relative improvement in reaction time and purportedly minimizes floor effects), mean accuracy rate, and speed-accuracy tradeoff score representing a multiplicative combination of speed and accuracy. The reaction time difference score failed to correlate significantly with any of their five higher cognition measures (r = .01 - .17), including the Raven's Advanced Progressive Matrices (Raven, 1941). In contrast, the ratio reaction time correlated significantly to four of the five (r = .19 - .48). Furthermore, task performance, as measured by either mean accuracy rate or with the speed-accuracy score, correlated strongly to their higher cognition (r = .33 - .62). The accuracy measures also showed age-related decline whereas the reaction time difference scores were smaller in older adults (suggesting that older individuals were faster), a puzzling result since age-related decline is expected in reaction time. Urry et al. argued that their results demonstrate that accuracy-based measures are more appropriate for measuring sequential learning than the existing reaction time methods. Their

results are in line with our own findings in other domains (e,g., attention control and task switching) and the reader should find them unsurprising given the discussion we have laid out regarding the problems with reaction time.

The Implicit Association Test

The Implicit Association Test (Greenwald et al., 1998), was designed to measure implicit biases, notably toward social groups, and is a widely used measure in social cognition research (Devine, Forscher, Austin, & Cox, 2012; Fridell, 2017; Jost et al., 2009; Lane, Banaji, Nosek, & Greenwald, 2007). The task requires subjects to categorize stimuli (e.g., words or images) into a category comprising two other elements, often a social group and a possible evaluation of that social group. For example, testing for racial bias would involve the subject judging whether words and pictures of faces are *white or good* or *black or bad* in one block, and whether those same stimuli are *black or good* or *white or bad* in a subsequent block. Implicit bias is inferred by subtracting reaction times from these blocks, and deviations from zero are assumed to reflect differential semantic or evaluative associations between the category components. For example, faster responses in *white or good* blocks than *black or good* blocks indicates a stronger association between "white" and "good" and an automatic preference for white individuals over black (Jost et al., 2009; Lane et al., 2007; but see Blanton & Jaccard, 2006).[15]

Policy makers and political researchers have suggested using the Implicit Association Test as an indicator of various socially relevant biases. It has been suggested as a screening tool for jury selection (Larson, 2010), for gauging the risk of convicted sexual predators' recidivism (Nunes, Firestone, & Baldwin, 2007), and for assessing racial biases in legislative decision

[15] In 2003, Greenwald, Nosek, & Banaji implemented a scoring procedure meant to control for speed-accuracy interactions. The method is now standard. Hence, at this time we do not regard such differences as significantly contributing to the Implicit Association Test's psychometric issues, but component scores remain highly correlated, contributing to reliability concerns.

making (Saujani, 2003). Blanton et al. (2009) claim it is psychology's most popular export to both the social sciences and the law. The degree to which the Implicit Association Test is utilized outside of basic psychological research is difficult to determine. However, the substantial amount of attention the test has garnered since its development along with the current political climate suggests that implicit bias testing could become more widespread. It is thus reasonable to speculate that implicit bias scores could be used to make judgments about individuals in high-stakes legal situations, and for selection of individuals for jobs and other positions.

While the constructs that the Implicit Association Test purport to measure are indeed pressing, some psychometric properties of the task raise concerns for its applicability *in individual differences contexts*. In a review, Lane et al. (2007) report test-retest reliabilities for different versions of the Implicit Association Test ranging from as low as .25 to a maximum of .69, with a median of .50. Gawronski, Morrison, Phills, and Galdi (2017) report moderately high internal consistencies for a race and a self-concept version of the task (Cronbach's alphas of .69 and .87, respectively), but also reported underwhelming test-retest reliabilities of .44 for the race and .63 for the self-concept Implicit Association Test.[16] These estimates are similar to many of the other difference score measures we have discussed so far, and, importantly, routinely fall below Nunnally's (1964) .80 guideline for basic research as well as his .95 guideline for high-stakes situations.

Noting these measurement difficulties, Cunningham, Preacher, & Banaji (2001) used a latent variable approach to measure performance on the Implicit Association Test because such methods isolate reliable variance and thus could be a way to circumvent the reliability issues of the Implicit Association Test at the task-level. Their assessment likewise indicated that Implicit

[16] The numerical value of reliability estimates for the Implicit Association Test vary depending on the attitudes being measured.

Association Test scores contain a large proportion of error variance. Moreover, scores on the task did not strongly relate to one another over time, even at the latent level.[17] In their analysis on 93 undergraduates, they found that only 46% of the reliable variance in their Implicit Association Test scores was stable across four testing sessions and that **scores on the first two administrations separated by only two weeks in time correlated at *r* = .31**. And, while latent variable analysis is better than simple correlational analysis and may be one way to avoid some of the reliability issues with differences scores (e.g., Gollwitzer et al., 2014), it requires a much larger sample size, longer administration time per subject, and is impractical in many situations (c.f., Friedman & Miyake, 2004). Further, those administering the test in applied scenarios may lack the requisite resources or knowledge of latent variable methodology to follow Cunningham et al.'s (2001) lead. Even if this was not the case, the reliability concerns of the Implicit Associations Test inspire little confidence in the veracity of results and conclusions about any particular individual's bias based on their scores.

In terms of validity, Blanton et al. (2009) argued that there is little-to-no empirical evidence that the scores on the Implicit Associations Test predict real-world behavior. According to them, empirical validation studies are few and far between, and studies which do claim to demonstrate validity at the individual level often have some combination of small sample sizes, outliers driving the effect, and/or conclusions that go beyond the reach of what their analyses and

[17] Some researchers may object that low stability/test-retest reliability is not an issue for tests that do not measure stable constructs, and that implicit bias is a context-sensitive construct that may fluctuate wildly over time and situations (e.g., Alkozei, Killgore, Smith, Dailey, Bajaj, & Haack, 2017; Gawronski et al., 2017). While we acknowledge this point, it does not absolve the Implicit Association Test of the measurement issues associated with difference score correlations. Further, this bolsters our stance that there needs to be more research and discussion before the test is deemed suitable as an individual differences measure for selection, placement, and legal settings. In other words, if implicit biases *are* reasonably stable, then higher test-retest reliability would be expected and the measure fails on this front. If implicit biases are not stable and do indeed fluctuate with time and depending on context, then much care needs to be taken in making high-impact decisions based on scores from a single administration.

methodology would permit. To that end, Blanton et al. (2009) reanalyzed data on the only two studies they could find that demonstrated the Implicit Associations Test's ability to predict workplace discrimination (in actual or simulated environments); they found one to be dependent on outliers and the other to have severe methodological flaws. In a more recent meta-analytic review, Oswald, Mitchell, Blanton, Jaccard, & Tetlock (2013) found that mean correlations between Implicit Association Test scores and criterion measures of ethnic and racial discrimination to be r = .12 - .15 in 46 published and unpublished studies. They noted that these results were lower than the more optimistic r = .20 - .24 reported by Greenwald, Poehlman, Uhlmann, and Banaji (2009) in their meta-analysis based on fewer studies and effects.[18]

To be clear, we are certainly not calling into question the existence of implicit biases per se (c.f., Jost et al., 2009), and a discussion of the theoretical and societal concerns with bias goes beyond the scope of this paper (see Blanton & Jaccard, 2006; Kang & Banaji, 2006). Furthermore, we are *not* advocating that researchers discontinue using the Implicit Association Test. The Implicit Association Test behaves in a theoretically predictable manner at the group level, making it a very useful tool for experimental purposes (e.g., Alkozei, et al., 2017; Blanton et al., 2009; Cunningham et al., 2001; Gawronski, 2002; Jost, 2018; Jost et al., 2009; Lane et al., 2007). Evidence also suggests that it is the single best measure of implicit attitudes currently available (Bar-Anan & Nosek, 2014). With that said, the reliability of the Implicit Association Test scores for <u>individual differences purposes</u> is dubious at best, and the correlations and effect sizes produced are generally small and fragile (Blanton et al., 2009; Oswald et al., 2013). The

[18] Note that low correlations involving the Implicit Associations Test can have multiple causes. For instance, poor reliability or specification of outcome measures would lead to low criterion validity of the Implicit Associations Test scores to no fault of the test itself. And though it is very possible that the outcome measures used to assess the validity of implicit bias measures may be problematic as well, we have to return to the consistent trend that all scores based on reaction time difference scores exhibit comparatively low correlations. As such, the simplest explanation is that the reliability and validity of the Implicit Associations Test raises concerns for its use in correlational settings.

severity of these measurement deficiencies raises serious doubt about the meaningfulness of individual scores on the Implicit Association Test, even if robust effects emerge at the experimental level. To put it another way, the Implicit Association Test is a great tool for showing the existence of implicit bias and to study how different contexts and manipulations affect its expression, but caution needs to be exercised in interpreting or making conclusions from any one individual's particular score, especially if that score is from a single administration. As such, there is a long way to go in validating implicit bias assessments such as the Implicit Association Test before it should be endorsed in high-stakes situations, and the suggestion that the Implicit Association Test be applied in legal procedures and other important areas should be considered with great caution.[19] Our recommendation is that the test not be used outside of experimental research on implicit biases until attempts to improve the Implicit Association Test's psychometric properties are proven successful (e.g., Nosek & Banaji, 2001). We suspect that this will involve reducing the test's reliance on reaction time difference scores as well as stricter adherence to existing best practices (e.g., basing interpretations on multiple administrations; A. G. Greenwald, personal communication, June 4, 2018).

The Attention Network Test

Whereas some propose that the Implicit Association Test has applications in various legal settings, the Attention Network Test has been widely implemented in clinical and developmental research. On the surface, much of this research has not been problematic from a psychometric standpoint because most studies utilizing this task are examinations at the group level, not the individual. However, the prevalence and implications of the Attention Network Test in

[19] Only a select few endorse applied applications of the Implicit Associations Test as most social psychologists likewise caution against such uses.

neuropsychological and developmental studies warrants closer examination as, here too, difference scores for individuals yield problematic interpretations.

The Attention Network Test combines the Eriksen flanker task and the Posner response time cueing task (Fan et al., 2002). Trials begin with a central fixation and then a cue conveying either, a) temporal information about when a series of arrows will appear, b) temporal information about when the arrows will appear *and* spatial information about whether a central target arrow will appear above or below the central fixation, or, c) no cues whatsoever. When the arrows appear, subjects must indicate whether the central arrow in a series points left or right. The direction of this central target can be either congruent or incongruent with the other arrows in the series.

The combination of cues yields three difference scores that have traditionally been interpreted as assessing three isolable attention networks: the orienting, the alerting, and the executive attention networks (Fan et al., 2002; Callejas et al., 2005; MacLeod, et al., 2010; Rueda et al., 2004). The scores for each network have widely disparate reliability estimates. A recent meta-analysis found Spearman-Brown corrected split-half estimates of .38 for the alerting network, .55 for the orienting network, and .81 for the executive network using reaction times from fifteen studies using the Attention Network Test (MacLeod et al., 2010). Another study administered two versions of the Attention Network Test across ten sessions and, according to a modified split-half estimate, executive network scores were reliable (.86) after including data from only two sessions. The alerting and orienting network scores required data from many more sessions before their reliability estimates reached significance (seven and ten sessions,

respectively; Ishigami & Klein, 2010).[20] While the reliability estimate for the executive network is larger than many other difference scores we have discussed, the task's overall reliability remains underwhelming. As Ishigami & Klein (2010) conclude, the Attention Network Test's reliability, "… is generally lower than is ideal for many purposes" (p. 127).

The children's version of the Attention Network Test is more concerning. Rueda et al. (2004) report split-half reliabilities of .59 for the executive network, .37 for the alerting network, and .02 for the orienting network. Ishigami & Klein (2015) administered the child Attention Network Test across ten sessions and, regardless of how many sessions they included, none of the network scores were reliable according to a modified split-half estimate. Furthermore, in many cases the network scores themselves were not statistically different from zero. Thus, the child version of the Attention Network Test is a poor candidate for a diagnostic tool or as a means of characterizing development longitudinally (e.g., Rueda et al., 2004; Suades-González et al., 2017). Ishigami & Klein (2015) suggested that the test be avoided altogether in research designs requiring multiple test administrations or when researchers are interested in correlating the network scores with another individual difference measure.

In addition to these reliability issues, the Attention Network Test illustrates another danger associated with difference score measures: false equivalence of component scores. Galvao-Carmona et al. (2014) contend that the cue conditions used to calculate the alerting and orienting network scores are not comparable in ways that researchers typically assume. To calculate the alerting network score, mean reaction times from a no-cue condition are subtracted

[20] In a version of the test developed by Callejas, Lupiañez, Funes, & Tudela (2005), an auditory rather than a visual cue conveys temporal information. Ishigami & Klein (2010) found that this version had superior psychometric properties to the original Attention Network Test, but was still problematic. We omit further discussion of this variant for brevity.

from a temporal cue condition.[21] This subtraction has been assumed to capture the difference between alert and not alert states, but Galvao-Carmona et al. (2014) suggest that the anticipation of an unknown event is a resource demanding process that slows responses in the no-cue condition. A subject in the no-cue condition is anticipating a stimulus but does not know whether that stimulus will require a response. Cued conditions remove this ambiguity: after a cue appears, the subject knows that the next stimulus presented will be a target that requires a response. The added demand in the no-cue condition means that the two component scores are difficult to compare and that it is not appropriate to analyze their difference. Similar objections pertain to the orienting network score. On temporally cued trials, subjects must monitor a much greater surface area for the appearance of the target. Upon the appearance of the target, they must also choose from a larger set of possible responses than on spatially cued trials. That is, they must either look up or down and then indicate whether the target points left or right (four possible responses) whereas they must only indicate whether targets point left or right on spatially cued trials (two possible responses).

The most significant contribution of the Attention Network Task to the psychological literature has been to demonstrate the relative independence of the three attention networks, as shown by a lack of significant correlations between the network scores (Fan et al., 2002; Ishigami & Klein, 2010). However, given the low reliability estimates of the alerting and orienting networks, and the impurity of difference score measures, a null correlation may not accurately reflect the attentional processes at play, but instead be due to psychometric

[21] In Fan et al. (2002), the alerting network score is calculated by subtracting a double-cue condition (cues simultaneously appearing above and below the fixation point where target arrows could appear) from the no-cue condition. The orienting network score is calculated by subtracting the spatial cuing condition from the central cuing condition (a cue replaces the fixation point). In Galvao-Carmona et al. (2014), no double-cue condition was used. We recognize this departure from established procedure, but still regard their analyses as illustrative.

shortcomings of the task (Galvao-Carmona et al., 2014; Miller & Ulrich, 2013; Redick & Engle, 2006). Thus, conclusions drawn from Attention Network Test scores regarding clinically relevant topics such as the efficacy of therapeutic and pharmacological interventions (e.g., Murphy & Alexopoulos, 2006), the character of different disorders (e.g., Pacheco-Unguetti, Acosta, Marqués, & Lupiañez, 2011; Urbanek et al., 2010), and developmental trajectories (Suades-Gonzáles et al., 2017) should be interpreted cautiously.

Alternatives to traditional reaction time and difference score measures

To this point, we have focused on the problems associated with reaction time and reaction time difference scores, with the previous section devoted to some areas within our field of study affected by said issues. However, in this section we are more optimistic and outline several statistical and methodological alternatives for differential, developmental, and applied practitioners to use. These alternatives will have varying degrees of usefulness depending on each individual researcher's specific goals, but researchers interested in individual and developmental differences with either children or older individuals ought to strongly consider these or other alternatives instead of either pure reaction time or reaction time difference scores.

Using component scores instead of difference scores

An understandable approach to combating reliability concerns of difference scores would be to use reaction time component scores (e.g., incongruent trials in the Stroop) instead. After all, simple reaction times are often highly reliable. Researchers have occasionally adopted this approach with tasks such as the Stroop and noted *some* increments in reliability and validity over difference scores (e.g., Kane et al., 2016; McVay & Kane, 2012).

Using component scores over difference scores may be an approach that works in some contexts. However, it should be done with caution. Pure reaction times are still highly sensitive

to speed-accuracy interactions, which would best be addressed with forethought during the planning of the experimental design. Another issue with using component scores is that they do not consider baseline performance and so task purity becomes a problem (e.g., Miller & Ulrich, 2013). Difference scores are used to isolate cognitive processes, combat the issue of task purity, and assess the efficacy of treatments or interventions. In some instances it would be ill-advised to use component scores instead. For instance, how would one assess post-treatment improvement with only a single, non-comparative, score? In the Stroop, it seems a great methodological undertaking to assess interference on incongruent trials relative to congruent trials without considering performance on both. In these instances, difference scores are used to better ensure the variance in the dependent variable is isolated to the process of interest. As such, using a component score instead of a difference score requires a reframing of the research question, employment of new measures or analyses, or concession that the test score is likely reflecting different processes.

Controls for the speed-accuracy tradeoff

As noted previously, speed-accuracy tradeoffs are a major concern with reaction time research, especially in differential settings in which respondents are likely to balance speed and accuracy differently, thereby introducing error variance and contaminating findings. There have been various attempts at accounting for, controlling, and directly studying speed-accuracy tradeoffs to minimize the potential negative impact it can have on reaction time data (see Heitz, 2014 for a more thorough review).

One of the first efforts to quantify the speed-accuracy tradeoff comes from a class of mathematical models known as random walk models (e.g., Fitts, 1966), with the diffusion model being the most widely applied of such models to cognitive tasks (Ratcliff, 1978; Ratcliff, Smith,

& McKoon, 2015). In these models, information is assumed to accumulate over time until the subject has enough confidence to make a response. Random walk models have parameters representing a subject's indecision time, rate of information accumulation, response bias, and response threshold. The response threshold parameter is the most relevant to speed-accuracy tradeoffs, because it represents precisely how much information needs to accumulate before a subject initiates a response, and thus is a representation of their speed-accuracy tendencies. But despite their appeal and ability to answer important questions about cognition (see Starns & Ratcliff, 2010, for one example), these models see limited use and have not been employed in large-scale correlational endeavors. This is because the models are complex and require sophisticated code to run (but see Wagenmakers, van der Maas, and Grassman, 2007), are often applicable to only simple two-choice tasks, and can require hundreds of trials to produce stable parameters depending on the application, with some researchers opting for trial numbers in the thousands (see Lerche, Voss, & Nagler, 2017). Nevertheless, the diffusion model is part of a broader field of research known as cognitive psychometrics. Cognitive psychometrics uses cognitive models for measurement purposes, and Wagenmakers et al. (2007) wrote that using such techniques sidesteps issues pertaining to speed-accuracy tradeoffs in cognitive research.

Another way to study speed-accuracy tradeoffs is through the implementation of experimental manipulations. These include, but are not limited to, verbal instructions (Howell & Kreidler, 1963), response deadlines (Pachella & Pew, 1968), and payoff matrices (Fitts, 1966). Instructions are less effective and thus less desirable (e.g., Heitz, 2014), but payoff matrices and deadlines can be effectively used to manipulate and study speed-accuracy tendencies. Response deadlines can also be used to force quick responding and thus see use in studies not specifically designed to assess speed-accuracy tradeoffs directly. The main problem with these experimental

manipulations, however, is that they often require advanced knowledge about the distributions of responses on the particular task. For example, the researcher must know where to set the deadline and how to structure the payoffs, and there is a level of arbitrariness involved. And, like diffusion modeling techniques, speed-accuracy manipulations often necessitate a larger number of trials. Nonetheless, these are highly effective manipulations with a rich history in reaction time research.

Finally, speed-accuracy tradeoffs can be assessed with specific analyses. The speed-accuracy tradeoff function, the conditional accuracy function, and the quantile-probability plot all depict speed-accuracy tradeoffs. Heitz (2014) describes these in detail, and so we will not. The appeal of these methods is that they can be done post-hoc, with the caveat that they work best in experiments containing multiple speed-accuracy conditions or manipulations (such as response deadlines or payoff matrices).

Integrative measures of reaction time and accuracy

Recently, researchers have suggested that one solution to speed-accuracy interactions is to meaningfully combine them into a single metric (e.g., Draheim et al., 2016; Hughes et al., 2014; Liesefeld, Fu, & Zimmer , 2015; Liesefeld & Janczyk, 2018; Vandierendonck, 2017; 2018). Some benefits of such measures include: 1) greater sensitivity to speed-accuracy interactions, 2) to the ability to measure performance on tasks traditionally measured using difference scores, 3) the researcher does not have to choose whether to use accuracy or reaction time as the dependent variable, 4) integrative measures contain more information than reaction time and accuracy separately, and 5) they can correct for differential speed-accuracy tradeoffs. Some drawbacks include: 1) debate over whether equal weighting of speed and accuracy is desirable and how to achieve it, 2) integration of speed and accuracy into a single metric will be

likely be arbitrary because researchers must decide to what extent speed versus accuracy contributes to the score, 3) some of the integrations require specific circumstances to be applicable, 4) data analysis becomes more complicated, and, 5) the resulting score might be difficult to interpret or even meaningless in raw form.

While integrative measures are not a recent development, a recent push to use them came when Hughes et al. (2014) identified the issues with using switch costs in task switching and assessed how three different integrative measures could improve the measurement of task switching. They analyzed task switching using traditional reaction time and accuracy switch costs, the rate residual score (number of correct responses per second; Woltz & Was, 2006), the inverse efficiency score (reaction time divided by accuracy rate; Townsend & Ashby, 1978), and their newly proposed binning procedure.[22] They found all three measures to be an improvement to either reaction time- or accuracy-based difference scores separately, and stated that the binning method was particularly promising.

Vandierendonck (2017) assessed seven integrative scoring techniques for cognitive data, including multiple binning procedures, the inverse efficiency score, the rate residual score, and his own linear integrated speed-accuracy score. Vandierendonck's primary concern was whether the integrative measures retained the information present in the individual reaction time and accuracy scores, and whether the integrative measures accounted for a larger proportion of variance than reaction time and accuracy considered separately. He argued against the Hughes et al. (2014) binning procedure due to its elaborate and arbitrary calculation, the tendency to

[22] In their binning procedure, the latency switch cost for each subject's *accurate* switch trial is rank-ordered from 1 to 10 (1 being the quickest) across all subjects such that each subject's accurate switch trial is assigned a corresponding bin value ranging from 1 – 10 indicating how quick or slow that trial was relative to both their baseline (repeat trial) reaction time and relative to the reaction times of other subjects. Then, inaccurate switch trials are assigned a bin value of 20. These values (1 – 10 or 20) are added up to produce a final bin score that represents overall task performance for that subject. See Hughes et al. (2014) or Vandierendonck (2017) for a more thorough explanation.

emphasize accuracy more heavily than reaction time, and a potential lack of independence between subjects (i.e., subject A's performance has an effect on subject B's scores).[23] He was most in favor of his own integrated measure but adjudged the rate residual score to be a very good measure and the inverse efficiency score to be trustworthy but worse than these other two. In a follow-up, Vandierendonck (2018) examined the rate residual, inverse efficiency, and his own integrative measure in a set of thirteen task switching experiments. He again determined that his own linear integrated speed-accuracy score was valid, but he qualified his previous stance on the other two, recommending that researchers avoid the rate residual score altogether and to only use the inverse efficiency score on data sets with low overall error rates.

In response to debate in the mental rotation literature of whether objects are represented holistically or as a collection of its parts, Liesefeld et al. (2015) argued that speed-accuracy tradeoffs explained conflicting results among researchers and why some subjects show effects of object complexity whereas others do not. They developed another integrative speed-accuracy measure called the balanced integration score to support their position. Liesefeld and Janczyk (2018) further tested this score along with the inverse efficiency score, rate-correct score, and Vandierendonck's (2017; 2018) linear-integrated speed-accuracy score.

As the name suggests, the appeal of the balanced integration score is that it equally weights reaction time and error rates by standardizing them into a z-score and taking the difference between the two. Liesefeld and Janczyk showed that the other integrative scores did not achieve an equal balance of speed and accuracy whereas their score, by definition, does. Further, the mean values of the other scores either increase or decrease with differing speed-

[23] Scores in the Hughes et al. (2014) binning procedure are not wholly independent because one step involves rank ordering responses across all respondents. It is therefore important to have a diverse and representative sample when applying it to a dataset, which is true of individual differences research in general.

accuracy levels. In other words, Liesefeld and Janczyk showed that their score was insensitive to differing levels of speed-accuracy tradeoffs whereas the other scores fluctuate depending on a respondent's location on the speed-accuracy tradeoff function. Whether insensitivity to speed-accuracy tradeoffs and equal weighting of speed and accuracy are desirable properties is a matter of discussion that goes beyond the scope of this paper. But it is noteworthy that the philosophy behind the balanced integration score is quite different than Vandierendonck's (2017; 2018) approach, which was instead to enlarge existing reaction time and accuracy effects in the data and preserve, not eliminate, speed-accuracy tradeoffs.

The problem with scores such as Vandierendonck's (2017; 2018) linear integrated measure and Liesefeld and Janczyk's (2018) balanced integrated measure is that they were designed to address speed-accuracy tradeoffs but not reliability concerns of differences scores. Liesefeld and Janczyk's balanced integration score is a difference score, and thus reliability and applicability in correlational research is a major concern just as it is with any other difference score. Further, the balanced integration score and Vandierendonck's linear integrated speed-accuracy score are within-contrast comparisons in that they are calculated separately for each type of trials. For instance, there is one score for incongruent trials in the Stroop and a separate score for congruent trials. How to combine these into a single score without using subtraction methodology, resulting in yet another difference score, is not clear. On the other hand, Hughes et al.'s (2014) binning procedure arrives at a single score that incorporates performance from both trial types and when considering both speed and accuracy. And, while the binning procedure still uses subtraction of the two trial types in the calculation, this calculation is done on a trial-by-trial basis, preserving some variability and leading to demonstrably higher reliability estimates than traditional difference scores (e.g., Draheim et al., 2016; Hughes et al., 2014). These qualities

make the binning procedure ideal for correlational analyses. As such, Liesefeld and Janczyk's and Vandierendonck's integrative measures are likely excellent tools for experimental researchers seeking to account for speed-accuracy interactions, but their utility as an alternative to difference scores for individual differences or developmental contexts remains to be shown.[24]

Accuracy-based measures

Another obvious solution to the problems of reaction time-based measures is to use raw accuracy or accuracy-based measures (capacity, *d'*, proportion correct, etc.), as Urry et al., (2015) suggested for measuring sequential learning. While accuracy-based **difference scores** usually have extremely low reliability (e.g., Hughes et al., 2014) and are typically of little use to individual differences researchers, *pure* accuracy measures are often very good for individual differences, provided that subjects make sufficient errors on the task. Accuracy-based measures thus require matching task difficulty with ability level of the tested population (see below for a discussion on adaptive measures).

It would be reasonable to assume that accuracy-based measures suffer the same pitfalls as reaction time measures in terms of ignoring speed-accuracy interactions. However, speed-accuracy interactions are more easily accounted for when using accuracy. Reaction time measures are problematic in part because they make it difficult for experimenters to influence participants' accuracy levels. Traditionally, participants are asked to respond as quickly and accurately as possible, an ambiguous instruction because increases in speed are made at the

[24] It has never been our position that the binning procedure proposed by Hughes et al. (2014) is a perfect integrative measure. Instead, in Draheim et al. (2016) we wanted to illuminate the problems with using reaction time and difference score measures to assess task switching and to determine whether *any* integrative measure could effectively be used to correct for these problems. In confirming this, we hoped more researchers would become aware of these techniques and continue to test and develop them, just as Vandierendonck (2017; 2018) and Liesefeld and Janczyk (2018) did. Furthermore, we only recommend using binning procedures in differential research, with a large sample size, a diverse population, and while also considering accuracy and reaction time independently. In other situations we recommend using another alternative.

expense of accuracy, and vice versa. This instruction also presupposes that subjects interpret instructions correctly, similar to one another, and that individuals have an equal ability to gauge the optimal speed-accuracy tradeoff. In contrast, many accuracy measures make reaction time more or less irrelevant by being auto-paced and/or by permitting the subject take as much time as needed to respond.

Antisaccade tasks are a good example of this and typically have desirable psychometric qualities. In a common version of this task (Kane, Bleckley, Conway, & Engle, 2001), the subject either properly inhibits their prepotent response to look toward the distractor (and can catch the target), or the distractor captures their attention and they miss the target. Since the target is presented for a fixed interval and their score is not contingent on responding within a given timeframe, subjects can take as long as they like to respond with minimal effect on accuracy. In that way, we describe this task as *reaction time-irrelevant*. Recall that Paap and Sawi (2016) designed a reaction time analog to the antisaccade and found that the resulting difference score had poor test-retest reliability, whereas performance in the accuracy version is highly reliable, correlates very strongly to other measures of executive functioning, and loads strongly onto an attention control factor (e.g., Kane et al., 2001; Shipstead et al., 2015; also see Table 1 in the Appendix). Paap and Sawi's finding is thus a rather convincing display of how the same paradigm can be reliable and valid when scored with accuracy but problematic when assessed using reaction time difference scores.

Other examples of reaction time-irrelevant measures that we use include fluid intelligence tests and complex span tasks. Subjects are permitted sufficient time to answer questions without feeling rushed as the time limit is imposed on the task as a whole, and not for individual items. Furthermore, these tasks begin with easy questions and progressively increase

in difficulty, and so subjects generally receive questions of difficulty that exceeds their ability level before time becomes a factor. Our complex span measures of working memory capacity have a response deadline for the processing (distractor) trials that is adaptive for each subject based on their reaction time on practice trials. In addition, the storage trials are auto-paced and task instructions specifically instruct the subject to take as long as necessary to respond to the recall screen at the end of each set – an improvement to the customary and contradictory "respond as quickly and accurately as possible" instructions.

Because speeded response are not as integral to performing well on the antisaccade, complex span, and fluid intelligence tasks, one component of the speed-accuracy issue is made mostly irrelevant. It is therefore not surprising that these accuracy-based measures consistently produce reliable dependent variables with high convergent and predictive validity. Accuracy rates in the fluid intelligence, working memory capacity, and antisaccade tasks we use can correlate to other measures as strongly as $r = .60$, whereas we rarely encounter reaction time-based measures with correlations exceeding $r = .30$ (and typically at or below $r = .20$). Such tasks also either control or account for reaction time such that speeded responses are not integral to performing the task. Accuracy-based measures can thus circumvent many of the issues with reaction time.

Signal detection theory

Signal detection theory is method with a rich history in experimental psychology (Green & Swets, 1966; Stanislaw & Todorov, 1999) in which patterns are deconstructed into two components – signal and noise. Detection theory is therefore used in studies in which subjects needs to discriminate between two types of stimuli. Like diffusion modeling, numerous parameters can be obtained using signal detection theory, and it is possible to disentangle speed

and accuracy. The most well-known parameter is the sensitivity index (*d'),* which is the mean separation of the noise vs. signal distributions for a subject.

In an unpublished guide to signal detection analysis for individual differences research, Paulhus and Petrusic (2010) noted some complications in using detection theory to assess individual differences, and that it had received only little attention among differential researchers except in a few notable exceptions – such as in personality research (Danziger & Larsen, 1989), clinical psychology debates concerning defense mechanisms, education research concerning the relationship between reading frequency and cognition, in the assessment of psychometrics of standardized tests, and fleetingly in memory research in the debate over distinct types of memory. Paulhus and Petrusic argued that the use of signal detection theory in these cases were contentious but often an improvement over existing methods. As such, there seems to be some potential in using signal detection and *d*' as an alternative to reaction time for assessing individual differences, but such an application has been employed only sparingly.

Adaptive or threshold tasks

Adaptive procedures are common in standardized testing (Way et al., 2010), and threshold procedures are frequently used in psychophysics (Leek, 2001). In general, these tasks involve administering different items or trials to subjects based on their individual performance such that the task becomes more or less challenging to better match the ability level of the respondent. Threshold tasks are a type of adaptive task that converge upon a specific value or score for the test-taker. For example, how similar two tones must be in frequency for the subject to respond at a 75% accuracy rate, or at what stimulus presentation rate the subject is 50% accurate.

There are numerous benefits to adaptive and threshold procedures. First, they have desirable psychometric properties because item difficulty is better matched to subject ability level, which is an improvement to administering the same trials to all subjects and taking an aggregate score. Because of this added precision, threshold tasks typically take less time to administer. Many traditional tasks can be modified to be adaptive and give more reliable estimates, including tasks that otherwise would rely on reaction time and/or difference scores. Finally, using adaptive procedures can serve as an effective control for either accuracy or reaction time, as they can be programmed to converge on a threshold at a certain accuracy rate. As mentioned in a previous section, our research team is currently exploring using threshold tasks as a replacement for the traditional Stroop and flanker tasks and early results are promising. Further, we have begun using adaptive sensory discrimination tasks to better understand the relationship between general discrimination ability and intelligence and to what extent attention control mediates this relationship (Tsukahara, Harrison, Draheim, Martin, & Engle, 2018).

Reliable components analysis

Noting the issues with simple difference scores, Caruso (2004) recommended using reliable component analysis instead. This method is a differential weighting technique that attempts to maximize the reliability of the resulting composite (see Cliff & Caruso, 1998). Caruso reanalyzed data from five cognitive assessment batteries using this procedure and reported reliabilities of .83 - .91 across 14 subtests traditionally measured using difference scores. These were significantly different from the reliabilities of the raw difference scores (ranging from .70 - .87). Caruso also noted, "The adequate reliability of difference scores found here contradicts conventional psychometric wisdom (e.g., Cattell, 1982; Cronbach & Furby, 1970; Lord & Novick, 1968) and certain empirical investigations (e.g., Malgady & Colon-

Malgady, 1991; Williams, Zimmerman, & Mazzagatti, 1987)" (p. 170). He attributed the surprisingly high reliability of the difference scores to him having selected the most popular assessment batteries for his analysis, and that these measures are likely popular in part due to their proven psychometric properties. Although the validity of scores from reliable components analysis needs to be established, it appears to be a procedure that can increase the reliability of difference scores without requiring researchers to resort to alternative tasks.

Residualized scores

Residualized scores involve rescaling performance on the outcome or post-test measure (in longitudinal research) or on the more demanding trial type (like incongruent Stroop trials). By regressing outcome or incongruent trial performance on baseline trial performance, the correlation between change and initial status is controlled for and only variability not explained by baseline performance is leftover (Castro-Schilo & Grimm, 2018). Residualized scores have thus gathered some attention as an alternative to difference scores. For instance, Steketee and Chambless (1992) championed their use in clinical psychology over simple gain scores; similarly Gollwitzer, Christ, and Lemmer (2014) argued for their use in social psychology; Williams, Zimmerman, and Mazzagatti (1987) found them to be more reliable than difference scores; and Kane et al. (2016) used residualized scores for some of their measures because they performed better than traditional difference scores. Despite this, reliabilities for Kane et al.'s measures were still suboptimal, as they report reliabilities as low as .25 with the rest (except one) of their residualized scores ranging from .48 - .59. Caruso (2004) also calculated the reliabilities of residualized scores and similar base-free difference scores in his assessment of cognitive batteries and reports that the residualized scores have higher reliability than simple difference scores. However, residualized scores were barely more reliable than simple difference scores

(and possibly not statistically significantly so, though no statistical test on the difference between the two scores was reported), whereas their reliable components analysis scores were much more reliable than both difference scores and residualized scores.

Although residualized scores may have *some* improved reliability over simple difference scores, these improvements are minor and residualized scores are still problematic in that they also follow subtraction methodology. Therefore, residualized scores are unlikely to perform markedly better than simple difference scores, and methods such as reliable components analysis appear more promising.

Polynomial regression

Polynomial regression is another noteworthy alternative because it sidesteps many of the reliability issues of difference scores while being applicable to tasks that are traditionally measured by said differences. It does so by using the intact components of a difference score to predict some variable of interest. Unlike other regression procedures, predictor values are entered into the model as linear terms but also as higher-order terms (i.e., x, x^2, x^3, etc. may all be included in the regression model) Since the method is not subtractive, reliable between-subjects variance is maintained while also preserving what difference scores aim to measure, the disparity between two theoretically interesting measures (Edwards, 2001; but see Tisak & Smith, 1994b). Moreover, polynomial regression allows for more robust analyses than do traditional difference scores. For example, Edwards (2001) argued that difference scores often implicitly contain untested hypotheses about the relationships between component scores, whereas polynomial regression allows such assumptions to be identified and tested. Further, difference scores are limited in the relationships and interactions between variables that they can convey because they are a single score. Maintaining the component scores allows polynomial regression to reveal

interactions between the component scores that may be of interest to researchers but that difference scores can obscure (Cohen, Nahum-Shani, & Doven, 2010; Edwards, 2001; Edwards & Parry, 1993).

Given these advantages, it may seem surprising that polynomial regression has not become a more popular method of measuring congruence. This is likely because difference scores do hold some noteworthy advantages, chiefly in terms of simplicity. Difference scores are at least interpretable, even if inferences made from them are suspect. On the other hand, even simple polynomial regression models contain a large number of coefficients, which are daunting for researchers to interpret – and this complexity has been a notable criticism of polynomial regression (Cohen et al, 2010; Edwards, 2001). Similar to diffusion modeling, complexity and unfamiliarity are likely major deterrents for many researchers (but see Edwards & Parry, 1993, for a method to make polynomial regression more palatable). Further, polynomial regression has seldom been used outside of organizational research and its implementation for studying individual differences in executive functioning would be novel. As such, the efficacy for these purposes is an empirical question that has not been answered. Additionally, as with all regression procedures, multicollinearity (interdependence among predictor variables) is a concern with polynomial regression. Multicollinearity can result in wildly inaccurate parameter estimates (Edwards, 2001) and biased/unstable standard errors leading to unstable p-values for predictors (Vatcheva, Lee, McCormick, & Rahbar, 2016). However, in research scenarios in which multicollinearity is not a concern, polynomial regression is a potentially fruitful alternative to difference scores.

Hierarchical Modeling

Rouder and Haaf (2018) recently provided their own perspective on the reason well-established experimental tasks are often poor individual differences measures. They argued that the problem lies with the *portability* – a basic property of classical test theory that underlying population values of a test are invariant to sample size (number of subjects *and* number of trials per task). They argued that that the use of aggregate scores violates portability and makes it such that reliability, correlations across tasks, and within-task effect sizes become functions of a researcher's sample size, resulting in different estimates of these values based on differences in samples. According to them, aggregating trial performance into a single aggregate score for a subject is problematic (violates portability) because trial-by-trial variation contaminates the aggregate score. They argued using hierarchical linear modeling of individual trial-level data can correct for this trial-by-trial noise and thus preserve portability (see their paper for a description of their model).

Rouder and Haaf (2018) focused on the problem with individual differences in attention control. Of particular interest was the well-established lack of correlation between reaction time interference effects in the color Stroop and arrow flanker tasks. They retested data from Hedge et al. (2018) to investigate whether this lack of correlation is due to substantive reasons, such as inhibition not being a unified construct (e.g., Rey-Mermet et al., 2018) or, as we have argued, methodological considerations, such as reliability and other measurement issues (e.g., Hedge et al., 2018). They found that, even with the use of hierarchical modeling to correct for trial-level variation, Stroop and flanker performance did not correlate. They interpreted this as substantive evidence against inhibition as a unified concept (similar to Rey-Mermet et al.), but they also stated statistical and methodological factors were still at play – specifically their suspicion that true individual variation may be much lower than believed.

Of note is that Rouder and Haaf (2018) still analyzed reaction time. Their concern was not with reaction time or difference scores as measures of performance, but rather trial-by-trial vs. aggregate-level data. If, as we argue, reaction time and difference scores are a problem, then even advanced modeling may not be of much benefit to researchers. However, as with polynomial regression, hierarchical modeling is a promising method, with more work needed to understand its usefulness in addressing issues with individual differences research.

Conclusions

It is not necessarily the case that reaction time paradigms which reliably show robust experimental effects will also produce reliable and valid individual differences. Although we agree that each researcher ought to assess the reliability and appropriateness of a dependent variable for their own purposes, it has been our experience that the cases in which reaction time difference scores are psychometrically justifiable given the available alternatives are the exception and not the norm. This is due to reaction time difference scores being necessarily less reliable than their component scores, often leading to demonstrably unreliable dependent variables. While we focus heavily on these issues as they relate to difference scores, pure reaction time measures are still problematic for their susceptibility to speed-accuracy interactions and because of the complexity of the processes underlying reaction time. These issues are magnified in individual differences and developmental studies, meaning that reaction time measures are particularly problematic in these settings.

The problems we outlined here are prevalent in many areas of scientific inquiry, both within and outside of psychology. Our position is that these problems have stifled theoretical advancements and led to both controversial and flawed conclusions in numerous fields of behavioral research. We strongly urge researchers to critically examine their own reaction time

variables and assess the extent to which these scores may be negatively influencing their own results and conclusions. Additionally, we urge anyone interested in basic individual differences research, developmental differences, clinical assessments, job selection, and other applied work to consider using alternative measures to avoid the methodological problems with reaction times, particularly reaction-time difference scores. A good alternative is accuracy-based measures in which reaction time is either controlled for or irrelevant. Another alternative is adaptive tasks which determine an individual's threshold for the ability or construct in question. We advocate for the continued use and exploration of integrative measures that combine speed and accuracy in a meaningful and, ideally, interpretable manner. Finally, we hope to see a rise in cognitive modeling for differential research, as such models have the potential to combat the known issues with using reaction time. A move away from reaction time in differential research is long overdue, and the scientific study of behavior will benefit from such a paradigm shift. At the very least, we hope that our efforts, along with the efforts of other researcher, to bring this discussion to a broader audience will lead to improved awareness and help facilitate discourse regarding how best to address these pressing issues.

References

Abutalebi, J., & Green, D. W. (2016). Neuroimaging of language controls in bilinguals: Neural adaptation and reserve. *Bilingualism: Language and Cognition, 19*(4), 689-698.

Alkozei, A., Killgore, W. D., Smith, R., Dailey, N. S., Bajaj, S., Haack, M. (2017). Chronic sleep restriction increases negative implicit attitudes toward Arab Muslims. *Scientific Reports, 7*(4285).

Allport, A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In: Umilta C., Moscovitch M. (Eds.), *Attention and Performance XV* (pp.421–452). Cambridge, Massachusetts: MIT Press.

Altman, E. M., & Gray, W. D. (2008). An integrated model of cognitive control in task switching. *Psychological Review, 115*(3), 602-639.

Baddeley, A. D. (1986). *Working Memory*. Oxford: Oxford University Press.

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.

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods, 46*(3), 668-688.

Bezruczko, N., Fatani, S. S., & Magari, N. (2016). Three tales of change: Ordinal scores, residualized gains, and rasch logits—When are they interchangeable?. *SAGE Open, 6*(3), 2158244016659905.

Bialystok, E. (1992). Selective attention in cognitive processing: The bilingual edge. In R. J. Harris (Ed.), *Cognitive processing in bilinguals* (pp. 501–513). Amsterdam: North-Holland.

Bialystok, E. (2017). The bilingual adaptation: How minds accommodate experience. *Psychological Bulletin, 143*(3), 233-262.

Bialystok, E., Craik, F. I. M., & Freedman, M. (2007). Bilingualism as a protection against the onset of symptoms of dementia. *Neuropsychologia, 45*(2), 459-464.

Bialystok, E., & Viswanathan, M. (2009). Components of executive control with advantages for bilingual children in two cultures. *Cognition, 112*(3), 494-500.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61*(1), 27-41.

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G, & Tetlock, P. E. (2009). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *Journal of Applied Psychology, 94*(3), 567-582.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*(4), 335-340.

Botwinick, J. & Storandt, M. (1973). Speed functions, vocabulary ability, and age. *Perceptual and Motor Skills, 36,* 1123-1128.

Brébion, G. (2001). Language processing, slowing, and speed/accuracy trade-off in the elderly. *Experimental Aging Research, 27*(2), 137-150.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296-322.

Cafri, G., Van Den Berg, P., & Brannick, M. T. (2010). What have the difference scores not been telling us? A critique of the use of self—ideal discrepancy in the assessment of body image and evaluation of an alternative data-analytic framework. *Assessment, 17*(3), 361-376.

Callejas, A., Lupiáñez, J., Funes, M. J., & Tudela, P. (2005). Modulations among the alerting, orienting, and executive control networks. *Experimental Brain Research, 167*(1), 27-37.

Campbell, D. T., & Kenny, D. A. (1999). A primer on regression artifacts. Guilford Publications.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Thousand Oaks, CA: SAGE.

Caruso, J. C. (2004). A comparison of the reliabilities of four types of difference scores for five cognitive assessment batteries. *European Journal of Psychological Assessment, 20*(3), 166-171.

Castro-Schilo, L., & Grimm, K. J. (2018). Using residualized change versus difference scores for longitudinal research. *Journal of Social and Personal Relationships, 35*(1), 32-58.

Cattell, R. B. (1982). *The inheritance of personality and ability: Research methods and findings*. New York, NY: Academic Press.

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

Chuderski, A. (2014). The relational integration task explains fluid reasoning above and beyond other working memory tasks. *Memory & Cognition, 42*, 448–463.

Chuderski, A. (2015). The broad factor of working memory is virtually isomorphic to fluid intelligence tested under time pressure. *Personality and Individual Differences, 85*, 98–104.

Cliff, N., & Caruso, J. C. (1998). Reliable component analysis through maximizing component reliability. *Psychological Methods, 3*, 291-308.

Cohen, A., Nahum-Shani, I., & Doveh, E. (2010). Further insight and additional inference methods for polynomial regression applied to the analysis of congruence. *Multivariate Behavioral Research, 45*(5), 828-852.

Collins, L. M. (1996). Is reliability obsolete? A commentary on "are simple gain scores obsolete?" *Applied Psychological Measurement, 20*, 289-292

Cook, P. B., & McReynolds, J. S. (1998). Lateral inhibition in the inner retina is important for spatial tuning of ganglion cells. *Nature Neuroscience, 1(8)*, 714.

Costa, A., Caramazza, A., & Sebastian-Galles, N. (2000). The cognate facilitation effect: Implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(5), 1283-1296.

Costa, A., Santesteban, M., & Ivanova, I. (2006). How do highly proficient bilinguals control their lexicalization process? Inhibitory and language-specific selection mechanisms are both functional. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(5), 1057.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671-684.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"—or should we?. *Psychological Bulletin, 74*(1), 68-80.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science, 12*(2), 163-170.

Danziger, P.R., & Larsen, J.D. (1989). Personality dimensions and memory as measured by signal detection. *Personality and Individual Differences, 10*, 809-811.

DeGutis, J., Wilmer, J., Mercado, R. J., & Cohan, S. (2013). Using regression to measure holistic face processing reveals a strong link with face recognition ability. *Cognition, 126*(1), 87-100.

DeVellis, R.F. (1991). *Scale development.* Newbury Park, NJ: Sage Publications

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology, 48*(6), 1267-1278.

Dinges, D. F., & Powell, J. W. (1985). Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments, & Computers, 17*(6), 652-655.

Donders, F. C. (1868/1969). Over de snelheid van psychische processen. [On the speed of mental processes.] (W. G. Koster, Trans.). In W. G. Koster (Ed.), Attention and performance II (pp. 412–431). Amsterdam: North Holland.

Donnelly, S., Brooks, P. J., & Homer, B. D. (2015). Examining the bilingual advantage on conflict resolution tasks: A meta-analysis. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.) *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, Austin, TX (596-601). Cognitive Science Society.

Draheim, C., Hicks, K. L., & Engle, R. W. (2016). Combining reaction time and accuracy: The relationship between working memory capacity and task-switching as a case example. *Perspectives on Psychological Science, 11*(1), 133-155.

Draheim, C., Martin, J. D., Tsukahara, J. S., Mashburn, C. A., & Engle, R. W. (2018). *Accuracy based tasks measure attention control better than reaction time-based ones : Evidence for the unity of inhibition*. Manuscript in preparation.

Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E. J. (2011). Testing theories of post-error slowing. *Attention, Perception, & Psychophysics, 74*(2), 454-465.

Edwards, J. R. (1994). The study of congruence in organizational behavior research: Critique and a proposed alternative. *Organizational Behavior and Human Decision Processes, 58*(1), 51-100.

Edwards, J. R. (2001). Ten difference score myths. *Organizational Research Methods, 4*(3), 265-287.

Edwards, J. R., & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal, 36*(6), 1577-1613.

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. (2018). Working memory and executive attention: A revisit. *Perspectives on Psychological Science, 13*(2), 190-193.

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

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 44, pp. 145-199). NY: Elsevier.

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

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience, 14*(3), 340-347.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, 201711978.

Fitts, P. M. (1966). Cognitive aspects of information processing: III. Set for speed versus accuracy. *Journal of Experimental Psychology, 71*(6), 849-857.

Forstmann, B.U., Tittgemeyer, M., Wagenmakers, E.-J., Derrfuss, J., Imperati, D., & Brown, S. (2011). The speed-accuracy tradeoff in the elderly brain: A structural model-based approach. *Journal of Neuroscience, 31*, 17242–17249.

Fridell L. A. (2017) The science of implicit bias and implications for policing. In *Producing bias-free policing: A science-based approach* (7-30). Springer.

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: a latent-variable analysis. *Journal of Experimental Psychology: General, 133(1),* 101.

Fuentes, L. J., & Campoy, G. (2008). The time course of alerting effect over orienting in the attention network test. *Experimental Brain Research, 185*, 667-672.

Galvao-Carmona, A., González-Rosa, J. J., Hidalgo-Muñoz, A. R., Páramo, D., Benitez, M.,

Izquierdo, G., & Vázquez-Marrufo, M. (2014). Disentangling the attention network test: Behavioral, event related potentials, and neural source analyses. *Frontiers in Human Neuroscience, 8*.

Gawronski, B. (2002). What does the Implicit Association Test measure? A test of the convergent and discriminant validity of prejudice-related IATs. *Experimental Psychology, 49*(3), 171-180.

Gawronski, B., Morrison, M., Phillis, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin, 43*(3), 300-312.

Gollwitzer, M., Christ, O., & Lemmer, G. (2014). Individual differences make a difference: On the use and the psychometric properties of difference scores in social psychology. *European Journal of Social Psychology, 44*(7), 673-682.

Gottman, J. M., & Rushe, R. H. (1993). The analysis of change: Issues, fallacies, and new ideas. *Journal of Consulting and Clinical Psychology, 61*(6), 907.

Gratton, G., Coles, M. G. H., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology: General, 121*(4), 480-506.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language, and Cognition, 1*(2), 67-81.

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(6),* 1464.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Attitudes and Social Cognition, 85*(2), 197-216.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17-41.

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

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*(3), 1166 1186.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience, 8*.

Heitz, R. P., and Engle, R. W. (2007). Focusing the spotlight: Individual differences in visual attention control. *Journal of Experimental Psychology: General, 136*, 217-240

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1992). Intraindividual change in text recall of the elderly. *Brain and Language, 42*(3), 248-269.

Hertzog, C., Vernon, M. C., & Rypma, B. (1993). Age differences in mental rotation task performance: The influence of speed/accuracy tradeoffs. *Journal of Gerontology, 48*(3), 150-156.

Howard, D. V., & Howard, J. H. (1992). Adult age differences in the rate of learning serial

patterns: Evidence from direct and indirect tests. *Psychology and Aging, 7*(2), 232-241.

Howell, W. C., & Kreidler, D. L. (1963). Information processing under contradictory instructional sets. *Journal of Experimental Psychology, 65*(1), 39-46.

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. *Behavioral Research Methods, 46*(3), 702-21.

Hultsch, D. F., MacDonald, S. W., & Dixon, R. A. (2002). Variability in reaction time performance of younger and older adults. The Journals of Gerontology Series B: *Psychological Sciences and Social Sciences, 57*(2), 101-115.

Hutton, S. B., & Ettinger, U. (2006). The antisaccade task as a research tool in psychopathology: a critical review. *Psychophysiology*, 43(3), 302-313.

Jersild, A. T. (1927). Mental set and shift. *Archives of Psychology, 89,* New York.

Johns, G. (1981). Difference score measures of organizational behavior variables: A critique. *Organizational Behavior and Human Performance, 27,* 443-463.

Jost, J. T. (2018). The IAT is dead, long live the IAT: Context-sensitive measures of implicit attitudes are indispensable to social and political psychology. *Current Directions in Psychological Science*, doi.org/10.1177/0963721418797309.

Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., & Hardin, C. D. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in Organizational Behavior, 29,* 39-69.

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., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(2), 336-358.

Kane, M. J., & Engle, R. W. (2003). Working memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, 132, 47-70.

Kane, M.J., Meier, M.E., Smeekens, B.A., Gross, G.M., Chun, C. A., Silvia, P.J., & Kwapil, T.R. (2016). Individual differences in the executive control of attention, memory, and thought, and their associations with schizotypy. *Journal of Experimental Psychology: General, 145*, 1017-1048.

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., Mackintosh, N. (2010). Implicit learning as an ability. *Cognition, 116,* 321-340.

Kessler, R. C. (1977). The use of change scores as criteria in longitudinal survey research. *Quality and Quantity, 11*(1), 43-66.

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.

Konrad, K., Neufang, S., Theil, C. M., Specht, K., Hanisch, C., Fan, J., … Fink, G. R. (2005). Development of attentional networks: An fMRI study with children and adults. *NeuroImage, 28*(2), 429-439.

Kroll, J. F., Dussias, P. E., Bice, K., & Perrotti, L. (2015). Bilingualism, mind, and brain. *Annual Review of Linguistics, 1*(1), 377-394.

Kroll, J. F., Gullifer, J. W., McClain, R., Rossi, E., & Martin, M. C. (2015). Selection and control in bilingual comprehension and production. In J. Schweiter (Ed.), *Cambridge Handbook of Bilingualism*. New York, NY: Cambridge University Press. 463-507.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods, 9*(2), 202-220.

Lane. L. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the Implicit Association Test: V. What we know (so far) about the method. In B. Wittenbrink & N. S. Schwartz (Eds.). *Implicit Measures of Attitudes: Procedures and Controversies*. New York: Guilford Press.

Larson, D. (2010). A fair and implicitly impartial jury: An argument for administering the Implicit Association Test during voir dire. *DePaul Journal for Social Justice, 3*(2), 139-172.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics, 63*(8), 1279-1292.

Lehtonen, M., Soveri, A., Laine, A., Järvenpää, J., de Bruin, A., & Antfolk, J. (2018). Is bilingualism associated with executive functioning in adults? A meta-analytic review. *Psychological Bulletin, 144*(4), 394-425.

Lerche, V., Voss, A., & Nagler, M. (2017). How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behavior Research Methods, 49*(2), 513-537.

Li, P., Legault, J., & Litcofsky, K. A. (2014). Neuroplasticity as a function of second language learning: anatomical changes in the human brain. *Cortex, 58*, 301-324.

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.

Liesefeld, H. R., Fu, X., & Zimmer, H. D. (2015). Fast and careless or careful and slow? Apparent holistic processing in mental rotation is explained by speed-accuracy trade-offs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(4), 1140.

Liesefeld, H. R., & Janczyk, M. (2018). Combining speed and accuracy to control for speed accuracy trade-offs(?). *Behavior Research Methods*, 1-21. https://doi.org/10.3758/s13428-018 1076-x.

Logan, G. D. (1985). Executive control of thought and action. *Acta Psychologica, 60*(2-3), 193 210.

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

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

Logie, R. H., Della Sala, S., Laiacona, M., Chalmers, P., & Wynn, V. (1996). Group aggregates and individual reliability: The case of verbal short-term memory. *Memory & Cognition, 24*(3), 305-321.

Lohman, D. F. (1989). Human intelligence: An introduction to advances in theory and research.

Review of Educational Research, 59(4), 333-373.

Lord, F. M. (1956). The measurement of growth. *ETS Research Bulletin Series, 1956*(1), i-22.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (pp.21-38). Madison, WI: University of Wisconsin Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

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

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390,* 279–281.

Luk, G., Bialystok, E., Craik, F. I. M., & Grady, C. L. (2011). Lifelong bilingualism maintains white matter integrity in older adults. *The Journal of Neuroscience, 31*(46), 16808 – 16813.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*(2), 163-203.

MacLeod, J. W., Lawrence, M. A., McConnell, M. M., Eskes, G. A., Klein, R. M., & Shore, D. I. (2010). Appraising the ANT: Psychometric and theoretical considerations of the Attention Network Test. *Neuropsychology, 24*(5), 637-651.

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.

McVay, J. C., & Kane, M. J. (2012). Drifting from slow to "D'oh!": working memory capacity

and mind wandering predict extreme reaction times and executive control errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(3), 525-549.

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology, Learning, Memory, and Cognition, 22*(6), 1423-1442.

Miller, J., & Ulrich, R. (2013). Mental chronometry and individual differences: Modeling reliabilities and correlations of reaction time means and effect sizes. *Psychonomic Bulletin Review, 20*, 819-858.

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*(1), 49-100.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences, 7*(3), 134-140.

Murphy, C. F., & Alexopoulos, G. S. (2006). Attention network dysfunction and treatment response of geriatric depression. *Journal of Clinical and Experimental Neuropsychology, 28*(1), 96-100.

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-Go association task. *Social Cognition, 19*, 625–666.

Nunes, K. L., Firestone, P, & Baldwin, M .W. (2007). Indirect assessment of cognitions of child sexual abusers with the Implicit Association Test. *Criminal Justice and Behavior, 34*(4), 454-475.

Nunnally, J. C. (1964). *Educational measurement and evaluation*. New York, NY:McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill, Inc.

Oberauer, K., Süß, H. M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence, 31* (2), 167-193.

Ollman, R. T. (1966). Fast guesses in choice reaction time. *Psychonomic Science, 6,* 155–156.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology, 105*(2), 171-192.

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin, 82*(1), 85-86.

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology, 66*, 232–258.

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

Paap, K. R., & Sawi, O. (2014). Bilingual advantages in executive functioning: Problems in convergent validity, discriminant validity, and the identification of theoretical constructs. *Frontiers in Psychology*, *5*(962).

Paap, K. R., & Sawi, O. (2016). The role of test-retest reliability in measuring individual and group differences in executive functioning. *Journal of Neuroscience Methods, 274*, 81-93.

Pacheco-Unguetti, A. P., Acosta, A., Marqués, E., & Lupiañez, J. (2011). Alterations of the attentional networks in patients with anxiety disorders. *Journal of Anxiety Disorders, 25*, 888-895.

Pachella, R. G., & Fisher, D. (1972). Hick's law and the speed-accuracy trade-off in absolute judgment. *Journal of Experimental Psychology, 92*(3), 378-384.

Pachella, R., & Pew, R. W. (1968). Speed-accuracy tradeoff in reaction time: Effect of discrete criterion times. *Journal of Experimental Psychology, 76*, 19–24.

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics, 44*, 369–378.

Paulhus, D. and Petrusic, W. M. (2010). *Measuring individual differences with signal detection analysis: A guide to indices based on knowledge ratings*. Unpublished manuscript

Peter, J. P., Churchill Jr., G. A., & Brown, T. J. (1993). Caution in the use of difference scores in consumer research. *Journal of Consumer Research, 19*(4), 655-662.

Pew, R. W. (1969). The speed-accuracy operating characteristic. *Acta Psychologica, 30*, 16-26.

Rabbitt, P. M. (1966). Errors and error-correction in choice-response tasks. *Journal of Experimental Psychology, 71*, 264–272.

Rabbitt, P. M. (1979). How old and young subjects monitor and control responses for accuracy and speed. *British Journal of Psychology, 70*, 305-311.

Rabbitt, P., Osman, P., Moore, B., & Stollery, B. (2001). There are stable individual differences in performance variability, both from moment to moment and from day to day. *The Quarterly Journal of Experimental Psychology: Section A, 54*(4), 981-1003.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59-108.

Ratcliff, R., Smith, P. L., & McKoon, G. (2015). Modeling regularities in response time and accuracy data with the diffusion model. *Current Directions in Psychological Science, 24(*6), 458-470.

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* (3), 164-171.

Redick, T. S., & Engle, R. W. (2006). Working memory capacity and Attention Network Test performance. *Applied Cognitive Psychology, 20,* 713-721.

Redick, T. S., Shipstead, Z., Meier, M. E., Montroy, J. J., Hicks, K. L., Unsworth, N., … & Engle, R. W. (2016). Cognitive predictors of a common multitasking ability: Contributions from working memory, attention control, and fluid intelligence. *Journal of Experimental Psychology: General, 145*(11), 1473-1492.

Regev, S., & Meiran, N. (2014). Post-error slowing is influenced by cognitive control demand. *Acta psychologica, 152*, 10-18.

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

Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology, 124*, 207-231.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological bulletin, 92*(3), 726.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability the difference score in the measurement of change. *Journal of Educational Measurement, 20*(4), 335-343.

Ross, D. A., Richler, J. J., & Gauthier, I. (2015). Reliability of composite-task measurements of holistic face processing. *Behavior Research Methods, 47*(3), 736-743.

Rouder, J. N., & Haaf, J. M. (2018). A Psychometrics of Individual Differences in Experimental Tasks. *PsyArxiv*. https://doi.org/10.31234/osf.io/f3h2k

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*, 763-797.

Salthouse, T. A. (1979). Adult age and the speed-accuracy trade-off. *Ergonomics*, *22*(7), 811-821.

Salthouse T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review 103*, 403– 428.

Salthouse, T. A., Fristoe, N., McGuthry, K. E., & Hambrick, D. Z. (1998). Relation of task switching to speed, age, and fluid intelligence. *Psychology and aging*, *13*(3), 445.

Saujani, R. M. (2003). Implicit Association Test: A measure of unconscious racism in legislative decision-making. *Michigan Journal of Race & Law, 8*(395), 395-423.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize?. *Journalof Research in Personality, 47*(5), 609-612.

Schouten, J. F., & Bekker, J. A. M. (1967). Reaction time and accuracy. *Acta Psychologica, 27*, 143-153.

Shipstead, Z., & Engle, R. W. (2013). Interference within the focus of attention: Working memory tasks reflect more than temporary maintenance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(1), 277-289.

Shipstead, Z., Harrison, T. L., & Engle. (2015). Working memory capacity and the scope and

control of attention. *Attention, Perception, & Psychophysics, 77*. 1863-1880.

Shipstead, Z., Harrison, T. L., & Engle. (2016). Working memory capacity and fluid intelligence: Maintenance and disengagement. *Perspectives on Psychological Science, 11*(6), 771-799.

Shipstead, Z., Lindsey, D. R., Marshall, R. L., & Engle, R. W. (2014). The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. *Journal of Memory and Language, 72,* 116-141.

Siegrist, M. (1997). Test-retest reliability of different versions of the Stroop test. *The Journal of Psychology, 131*(3), 299-306.

Simon, J. R., & Rudell, A. P. (1967). Auditory SR compatibility: the effect of an irrelevant cue on information processing. *Journal of Applied Psychology, 51*(3), 300-304.

Spearman, C. C. (1910). Correlations calculated from faulty data. *British Journal of Psychology, 3*, 271-295.

Sperber, R. D., McCauley, C., Ragain, R. D., & Weil, C. M. (1979). Semantic priming effects on picture and word processing. *Memory & Cognition, 7*(5), 339-345.

Sriram, N., Greenwald, A. G., & Nosek, B. A. (2010). Correlational biases in mean response latency differences. *Statistical Methodology, 7*(3), 277-291.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137-149.

Starns, J. J., & Ratcliff, R. (2010). The effects of aging on the speed-accuracy compromise: Boundary optimality and the diffusion model. *Psychology and Aging, 25*(2), 377-390.

Steketee, G., & Chambless, D. L. (1992). Methodological issues in prediction of treatment outcome. *Clinical Psychology Review, 12(*4), 387-400.

Streiner, D L., and Norman, G. R. (1995). *Measurement scales: A practical guide to their development and use (2nd ed.).* Oxford: Oxford University Press.

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

Thomas, D. R., & Zumbo, B. D. (2012). Difference scores from the point of view of reliability and repeated-measures ANOVA: In defense of difference scores for data analysis. *Educational and Psychological Measurement, 72*, 37-43.

Tisak, J., & Smith, C. S. (1994a). Defending and extending difference score methods. *Journal of Management, 20*(3), 675-682.

Tisak, J., & Smith, C. S. (1994b). Rejoinder to Edwards's comments. *Journal of Management, 20*, 691-694.

Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan and F. Restle (Eds.), *Cognitive Theory Vol. III* (pp. 200 239). Hillsdale, NJ: Erlbaum Associates.

Trafimow, D. (2015). A defense against the alleged unreliability of difference scores. *Cogent Mathematics, 2*, 1064626.

Tsukahara, J. S., Harrison, T. L., Draheim, C. D., Martin, J. D., & Engle, R. W. (2018) *Attention control as a mediator of the sensory discrimination and intelligence relationship.* Manuscript submitted for publication.

Unsworth, N., & Engle, R. W. (2005). Individual differences in working memory capacity and learning: Evidence from the serial reaction time task. *Memory & Cognition, 33*(2), 213 220.

Unsworth, N., Redick, T. S., Spillers, G. J., & Brewer, G. A. (2012). Variation in working memory capacity and cognitive control: Goal maintenance and microadjustments of control. *The Quarterly Journal of Experimental Psychology, 65*(2), 326-355.

Unsworth, N., & Robison, M. K. (2016). Pupillary correlates of lapses of sustained attention. *Cognitive, Affective, & Behavioral Neuroscience, 16*(4), 601-615.

Unsworth, N., & Spillers, G. J. (2010). Working memory capacity: Attention control, secondary memory, or both? A direct test of the dual-component model. *Journal of Memory and Language, 62*(4), 392-406.

Urbanek, C., Weinges-Evers, N., Bellmann-Strobi, J., Bock, M., Dörr, J., Hahn, E., … Paul, F. (2010). Attention Network Test reveals alerting network dysfunction in multiple sclerosis. *Multiple Sclerosis, 16*(1), 93-99.

Urry, K., Burns, N. R., & Baetu, I. (2015). Accuracy-based measures provide a better measure of sequence learning than reaction time-based measures. *Frontiers in Psychology, 6*.

Vatcheva, K. P., Lee, M., McCormick, J. B., & Rahbar, M. H. (2016). Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology (Sunnyvale, Calif.), 6*(2).

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavioral Research Methods, 49*(2), 653-673.

Vandierendonck, A. (2018). Further tests of the utility of integrated speed-accuracy measures in task switching. *Journal of Cognition, 1*(1), 1-16.

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

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

Way, W. D., Twing, J. S., Camara, W. J., Sweeney, K., Lazer, S. & Mazzeo, J. (2010). *Some considerations related to the use of adaptive testing for the Common Core Assessments*. Retrieved from http://www.ets.org/research/policy_research_reports/publications/paper/2010/icfi.

Weinreich, U. (1953). Languages in contact, findings and problems. New York, NY: Linguistic Circle of New York.

Whitehead, P. S., Brewer, G. A., & Blais, C. (July 26, 2018). Are cognitive control processes reliable? *Journal of experimental psychology. Learning, memory, and cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0000632

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

Williams, R. H., Zimmerman, D. W., & Mazzagatti, R. D. (1987). Large sample estimates of the reliability of simple, residualized, and base-free gain scores. *The Journal of Experimental Education, 55*(2), 116-118.

Woltz, D. J., & Was, C. A. (2006). Availability of related long-term memory during and after attention focus in working memory. *Memory & Cognition, 34*(3), 668-684.

Wöstmann, N. M., Aichert, D. S., Costa, A., Rubia, K., Möller, H. J., & Ettinger, U. (2013). Reliability and plasticity of response inhibition and interference control. *Brain & Cognition, 81,* 82-94.

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement, 19*(2), 149-154.

Appendix

<Table 1 >

<Table 2>

<Table 3>