EXECUTIVE FUNCTION IN DOWN SYNDROME:

A META-ANALYSIS

by

ANDREW TUNGATE

FRANCES CONNERS, COMMITTEE CHAIR
ANSLEY GILPIN
MATTHEW JARRETT
JEFFREY PARKER
DEBORAH CASPER

A DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
the Department of Psychology
in the Graduate School of
The University of Alabama

TUSCALOOSA, ALABAMA

2018
ABSTRACT

Executive function refers to a set of cognitive processes involved in goal-oriented behavior—especially inhibition, attention shifting, and working memory. In Down syndrome, there was an established weakness in executive function but the exact nature of the weakness was not well established. Executive functions are associated with a host of important outcomes, so it is important to understand executive function ability in Down syndrome. The current meta-analysis included 57 studies that compared a group with Down syndrome to a typically developing mental age matched group on at least one executive function task. Overall, individuals with Down syndrome performed significantly poorer on EF tasks than their typically developing mental age matched peers. Inhibition was a relative strength in the executive function profile, but it is still a weakness compared to the matched groups. There was a medium to large effect size across all executive functions but effect sizes may have been restricted due to skew. Implications for interventions and future research are discussed.
DEDICATION

This dissertation is dedicated to individuals with Down syndrome, their families, and all who have the opportunity to meet them.
LIST OF ABBREVIATIONS AND SYMBOLS

α  Chronbach’s alpha; referring to inter rater agreement

d  Cohen’s d

k  Number of effect sizes

M  Arithmetic mean

p  Probability associated with the occurrence under the null hypothesis of a value as extreme as or more extreme than the observed value

r  Pearson product-moment correlation

t  Computed value of a t test

<  Less than

=  Equal to
ACKNOWLEDGEMENTS

I am grateful for the opportunity to thank all colleagues, friends, family, and faculty members who helped me complete this project. Camilla, I could not have done this dissertation without you. Thank you for supporting me through this process, helping code for reliability, and for your unending patience while I worked on this project. I frequently talked to you about this paper and no wife should ever have to hear that much about a meta-analysis. Eloise, thank you for being a constant reminder of joy through this process.

Fran, I continue to be indebted to you for all that you have done for me. I could not have done this without your thoughtful guidance, insight, patience, and support during the last several years. I would not be half the scientist I am today without your direction. I would also like to thank my committee members—Dr. Ansley Gilpin, Dr. Matthew Jarrett, Dr. Jeffrey Parker, and Dr. Deborah Casper. Your support for me is very appreciated the insight you provided me on this project has been invaluable—thank you.
# CONTENTS

ABSTRACT ......................................................................................................................... ii
DEDICATION .................................................................................................................... iii
LIST OF ABBREVIATIONS AND SYMBOLS ....................................................................... iv
ACKNOWLEDGEMENTS ..................................................................................................... v
LIST OF TABLES ................................................................................................................ vii
LIST OF FIGURES ............................................................................................................. viii

1. INTRODUCTION ......................................................................................................... 1
   Down Syndrome ........................................................................................................ 2
   Executive Function .................................................................................................... 4
   Hypotheses ................................................................................................................ 26

2. METHODOLOGY ....................................................................................................... 28
   Data Collection ......................................................................................................... 28
   Inclusion Criteria ..................................................................................................... 29
   Coding ...................................................................................................................... 32
   Analysis Plan ................................................................................................ .......... 37

3. RESULTS ................................................................................................................... 45

4. DISCUSSION ............................................................................................................. 55

REFERENCES ............................................................................................................... 64
# LIST OF TABLES

1. Task Selection Criteria ........................................................................................................ 32
2. Executive Function Subgroup Comparisons .......................................................................... 47
3. Moderator Analyses for Overall Composites ......................................................................... 48
4. Moderator Analyses for EF Subscales ..................................................................................... 50
5. Mean Low and High Skew, and Percentage of Skewed Scores ............................................ 52
LIST OF FIGURES

1. Flow Chart for Search Process .................................................................................. 31
2. Example Funnel Plot .................................................................................................. 42
3. Composite EF Forrest Plot ......................................................................................... 46
1. INTRODUCTION

Down syndrome (DS) is a genetic disorder most commonly caused by complete trisomy of chromosome 21 (Lejeune, Turpin, & Gautier, 1959). It is the most identifiable genetic cause of intellectual disability (ID; Bittles, Bower, Hussain, & Glasson, 2007)—occurring in approximately 1 out of every 700 live births in the United States (Parker et al., 2010). The term executive function (EF) refers to a set of cognitive abilities used during goal-oriented behavior. Executive function is often referred to as “executive control,” which highlights how the components of EF are well thought of as control mechanisms (Miyake et al., 2000). There is not a definitive list of the components of EF (Banich, 2009). However, there is agreement that three components are fundamental to EF: inhibition (deliberate suppression of a prepotent response), shifting (ability to switch between tasks or rules efficiently), and the updating of working memory (WM; monitoring, updating, and manipulation of information in the mind; Diamond, 2013).

Executive function is associated with a host of important outcomes, ranging from academic achievement in school-aged children (Best, Miller, & Naglieri, 2011), to independent activity levels in older adults (Cahn-Weiner, Boyle, & Malloy, 2002). Especially relevant to DS are its association with depression (Tavares et al., 2007), school readiness (Morrison, Ponitz, & McClelland, 2010), food intake (Riggs, Spruijt-Metz, Sakuma, Chou, & Pentz, 2010), physical health (Miller, Barnes, & Beaver, 2011), and quality of life (Brown, & Landgraf, 2010). Importantly, EF is amenable to interventions—especially in individuals with poorer EF (see Diamond & Lee, 2011). For example, Davis et al. (2011) conducted a study that investigated the
effect of exercise on EF in overweight 7–11 year-olds. The children were randomly assigned to a control group, low-dose exercise group, or a high-dose exercise group for approximately 13 weeks. They reported improvements in a higher-level EF and math achievement for the high-dose exercise group, but there were no significant improvements in the low-dose exercise and control groups.

Down syndrome has a specific cognitive phenotype including several cognitive weaknesses (Silverman, 2007). General EF is considered one of the weaknesses in DS (Daunhauer et al., 2014; Lanfranchi, Jerman, Dal Pont, Alberti, & Vianello, 2010; Rowe, Lavender, & Turk, 2006). More specifically, Daunhauer et al. (2014) reported that individuals with DS had significantly poorer EF than their typically developing mental-age (TDMA) matched peers in WM, inhibition, and planning but not in shifting. Daunhauer et al.’s findings regarding shifting are consistent with Lee et al.’s (2011) findings, but not Lanfranchi et al. (2010). Lanfranchi et al. (2010) reported that there had been conflicting results on the exact profile of the EF weakness in DS. Nobody has employed a meta-analytic review to establish the EF profile in DS. The purpose of the current paper is to further elucidate the profile of EF in DS by conducting a meta-analytic review.

**Down Syndrome**

As mentioned previously, DS is the most identifiable genetic form of ID (Bittles et al., 2007). Bull (2011) reported that approximately 95% of cases result from a free extra copy of chromosome 21 (i.e., trisomy of chromosome 21). They estimated 3%–4% of cases result from an unbalanced translocation. More specifically, there is an extra copy of chromosome 21, but it is attached to another chromosome, often chromosome 14. The remaining 1%–2% of cases result from mosaicism, which is a mixture of normal cells and cells with trisomy of chromosome 21.
Physically, individuals with DS have several features that make them distinguishable from typically developing (TD) individuals. Common features in DS include a short stature, flat face, microcephaly, hypotonia, and broad, short hands (Korenberg et al., 1994).

In addition to distinct physical characteristics in DS, there are several health conditions associated with its phenotype. Namely, hearing loss (Dahle & McCollister, 1986); vision problems (John, Bromham, Woodhouse, & Candy, 2004); congenital heart defects (Freeman et al., 1998); obesity (Rubin, Rimmer, Chicoine, Braddock, & McGuire, 1998); obstructive sleep apnea (Marcus et al., 1991); certain cancers (Hill et al., 2003); hypothyroidism (Roizen & Patterson, 2003); and Alzheimer’s disease (Zigman et al., 2008). A few of the characteristic health conditions in DS are associated with EF. For example, obesity in children and adolescents is associated with poorer performance on various tasks assessing EF (see Barkin, 2013, for review). Chen, Spanò, and Edgin (2013) conducted a study on the relationship between EF and sleep in DS. They found that participants who had more sleep disturbances performed poorer on EF measures. Last, a decline in EF is related to early decline in Alzheimer’s disease (e.g., Baudic et al., 2006).

The cognitive phenotype in DS is characterized by several weaknesses and a few relative strengths. Aside from a weakness in EF, there are weaknesses in language (Næss, Lyster, Hulme, & Melby-Lervåg, M., 2011), especially in expressive language (Abbeduto et al., 2001); verbal short-term memory (Jarrold, Baddeley, & Hewes, 2000); wayfinding ability (Davis, Merrill, Conners, & Roskos, 2014); and verbal WM (Silverman, 2007). They may have a relative strength in receptive vocabulary compared to TDMA matched groups (Næss et al., 2011). Adding to difficulties later in life are several well-established links between DS and Alzheimer’s disease (Lott & Head, 2001). For example, there are increased levels of the plaque associated
with Alzheimer’s in the brain of individuals with DS for the duration of their lives (Zigman & Lott, 2007).

**Executive Function**

**History.** Executive Function refers to a set of control mechanisms, typically used in novel situations. Executive function research emerged from research on patients with frontal lobe damage. One of the most famous cases of frontal lobe damage occurred in 1848 to Phineas Gage. Phineas Gage suffered an accident that damaged his left frontal lobe and much of the nearby white matter (Van Horn et al., 2012). The case of Phineas Gage sparked curiosity in the field and started a thorough investigation into the function of the frontal lobe—an investigation that continues today.

Specific areas of the frontal lobe have been associated with many of the components of EF (e.g., response inhibition, attention shift, WM, etc.; Dias, Robins, & Roberts, 1996; Duncan & Owen, 2000). However, EFs are not just localized to the frontal lobe. There is evidence that EFs may be monitored by the frontal lobes, but there is a complex network of areas that are utilized by EF (Alvarez & Emory, 2006; Elliot, 2003). Although the concept “executive function” emerged from patients with frontal lobe damage, we now know EF is more complex. Accordingly, several different theories have been proposed to explain EF (Banich, 2009). Only important theoretical issues relevant to the current paper and its interpretation are discussed here—see the introduction of Miyake et al. (2000) for a more thorough review.

**Theory.** Teuber (1972) was the first to raise the question of whether frontal lobe function acted as one underlying mechanism or whether EFs were distinct. For several decades, researchers provided seemingly conflicting evidence on the unity and diversity of EF and consensus remained elusive for many years. More recently, the field has gravitated toward an
integrative approach—the idea that EFs, in adults, are both unitary and distinct (see Miyake & Friedman, 2012 for a short review). To put it another way, the components of EF have both shared (unitary) and unique (distinct) variance. For children, the unity and diversity of EF is more complex.

**Unity and diversity of executive function in pre-school children.** Is EF unitary early in development and, if so, does it differentiate as children get older? It is important to study the development of EF in TD children because individuals with DS are more similar in ability to TD children than TD adults. In pre-school children, there is evidence supporting both a one- (Wiebe, Espy, & Charak, 2008; Wiebe et al., 2011) and a two-factor model of EF (Miller, Giesbrecht, Müller, McInerney, & Kerns, 2012). Miller et al.’s (2012) two-factor model retained inhibition and WM in the model. There is evidence that EFs, even in the preschool years, become more distinct with age. Lerner and Lonigan (2014) conducted a study in 289 children, 45 to 63 months old that evaluated the relationship between WM and inhibitory control. Using Confirmatory Factor Analysis (CFA), they reported that a two-factor model fits the data better than a one-factor model. Consistent with findings from Tsujimoto, Kuwajima, and Sawaguchi (2007), Lerner and Lonigan (2014) noted that age influenced the correlation between factors. That is, inhibition and WM were more strongly correlated in younger children than in older children in their sample. Because there are a small number of studies on the structure of EF in preschool children, I cannot draw strong conclusions. However, the current evidence is consistent that EF ability in preschool is more unitary than it is in school-aged children. In sum, there is both unity and diversity in EFs in children as young as 3–5, but it is less distinct the younger the children are.
Unity and diversity of executive function in school-aged children. In school-aged children, there is evidence supporting both a two- (Huizinga, Dolan, & van der Molen, 2006; van der Sluis, de Jong, & van der Leij, 2007) and a three-factor structure of EF (Lehto, Juujärvi, Kooistra, & Thekinen, 2003; Rose, Feldman, & Jankowski, 2011). Both studies supporting the two-factor structure reported the best fit with shifting and WM in the model, but not inhibition. The studies supporting a three-factor model had inhibition, shifting, and updating of WM (Lehto et al., 2003; Rose et al., 2011). Shing, Lindenberger, Diamond, Li, and Davidson (2010) investigated inhibition and updating of WM in children ages 4–14. They found that EFs in older children (ages 9.5–14.5) were best fit by a two-factor structure; whereas, a one-factor model was more appropriate for younger children (ages 4–9.5). Shing et al.’s study is one of many that highlights that EFs become more distinct as children get older (see also Brydges, Fox, Reid, & Anderson, 2014; Huizinga, Dolan, & van der Molen, 2006; and Lee, Bull, & Ho, 2013). The time at which EFs become distinct is a topic of great interest for obvious reasons. Primarily because EF is associated with a host of important outcomes, so it is important to understand critical times in its development.

Hot and cool executive function. Zelazo and Müller (2002) introduced the concept of hot and cool EF as a distinction from the general-domain view of EF. More specifically, they made a distinction between affective aspects of EF associated with the orbitofrontal cortex (hot EF) and cognitive aspects of EF associated with the dorsolateral prefrontal cortex (cool EF). In plain English, hot EF is associated with affective aspects of EF, and cool EF is associated with cognitive aspects of EF. Zelazo and Müller (2002) point out that the cool aspects of EF are studied more often than the hot aspects of EF, and they argue that both warrant research because
hot EF is used in situations with high affective involvement by appraising the affective significance of the situation—something cool EF is unable to do.

An example of a hot EF task would be a delay of gratification task. In children, a delay of gratification task often requires the participant to choose whether they want a small award right away (e.g., their favorite snack), or a larger award (e.g., more of their favorite snack) at a future time. Delay of gratification exemplifies hot EF because the rewards have emotional significance to the participants. An example of a cool EF task is a card-sorting task like the Wisconsin Card Sort Task (WCST). The WCST presents participants with cards that have various dimensions (e.g., different colors, patterns, and shapes) and, as the experimenter presents cards, the participant needs to identify and follow the unspoken rule the cards are being sorted by. As the experimenter presents more cards, the participants are required to sort the cards, and they are told whether they sorted correctly. Once participants achieve a certain number correct consecutively, the experimenter changes the rule (i.e., the dimension) the cards are sorted on, and the participants must identify the new rule. The WCST exemplifies cool EF because it requires participants to shift flexibly between rules, but it has no affective component—it is a purely cognitive task.

Research on hot and cool EF is primarily done in children—being a developmentally early distinction in the “general-domain” view of EF—and is a framework that has been adopted by several researchers. In DS, Lee et al. (2015) adopted the framework and measured EF using a proxy-report. A proxy report is a questionnaire—in this case, on EF—that is filled out by a caregiver or teacher about the participant. They reported that individuals with DS have more difficulties with cool EF than with hot EF. Lee et al.’s (2015) findings are consistent with previous work in DS (e.g., Daunhauer et al., 2014; Lee et al., 2011). In TD preschoolers,
Hongwanishkul, Happaney, Lee, and Zelazo (2005) studied hot and cool EF. They reported that tasks assessing hot and cool EF were distinct from each other, and they were unique in their relationship to general intelligence. They concluded that this was evidence supporting hot and cool EF as concepts, and that that the two concepts are distinct from one another. However, not everyone agrees that hot and cool EF are distinct.

The distinction of hot and cool EF—or a distinction in affective versus cognitive aspects of EF—is one of the important debates in EF research. There is evidence that contradicts Hongwanishkul et al.’s (2005) conclusions for affective inhibition (e.g., delay of gratification) versus cognitive inhibition (e.g., a Stroop task). In a CFA, affective and cognitive inhibition were highly related and loaded onto the same factor (Friedman & Miyake, 2004). Further, Diamond (2016) highlighted research that demonstrated performance on an affective inhibition task was worse when preceded by a cognitive inhibition task rather than a difficult cognitive task not tapping inhibition (see Muraven & Baumeister, 2000). Muraven and Baumeister’s (2000) results suggest inhibition relies on a single resource. Last, Cohen, Berkman, and Lieberman (2013) conducted a review of neural imaging studies and concluded that cognitive, behavioral, and affective aspects of self-control operate on the same neural substrates (i.e., the right ventrolateral prefrontal cortex). Although the right ventrolateral prefrontal cortex refers to a broad area of the brain, their review casts doubt that inhibitory control can be parsed into hot and cool EF on the basis of the neural substrates.

In sum, hot and cool EF has provided a useful framework for researchers studying EF, but the distinction may not be necessary or warranted. There is not enough evidence to completely discredit the concept of hot and cool EF. However, there is evidence that casts doubt on the theoretical foundation of hot and cool EF. For that reason, I do not discuss the findings of
the current paper in the context of the “hot and cool” framework. Rather, I discuss EF using the integrative framework proposed by Miyake et al. (2000).

**Outcomes associated with executive function.** As mentioned previously, EF is associated with a host of positive and negative outcomes—highlighting the need to establish the EF profile in DS. Especially important, EF is associated with school readiness and academic achievement, successful transition to adulthood, job success, and social, physical, and mental health.

**School readiness and academic achievement.** Several studies have reported that EF is critical for school readiness in TD children (Diamond, 2014), and children with Autism Spectrum Disorder (Pellicano, Kenny, Brede, Klaric, Lichwa, & McMillin, 2017). School readiness is measured several different ways (High, 2008). For example, it has been measured by assessing language skills, literacy and math skills, attention ability, general knowledge, and emotional development. It is intuitive that EF’s—the cognitive processes associated with successful goal-oriented behavior—would be associated with school readiness. However, it is important to note that there is still significant relationship between EF tasks and school readiness after controlling for general intelligence (e.g., Blair & Razza 2007; Bull & Scerif, 2001; Normandeau & Guay, 1998; Willoughby, Magnus, Vernon-Feagans, Blair, & Family Life Project Investigators, 2017). In the classroom, a student that can inhibit task-unrelated behaviors—a characteristic of better EF—is considered better prepared for school than students that are less able to do so. Although the student above benefits from their general intelligence, EFs account for additional school readiness on top of general intelligence.

There are similar associations between EF and academic achievement. More specifically, WM and inhibition are especially related to academic achievement (see Best et al., 2011) whereas the relationship with shifting is less consistent. Behaviorally, school-aged children
benefit from the same EF skills that aid school readiness (e.g., the ability to inhibit distracting thoughts, leaving their seat, etc.). However, related to academic achievement, better EFs aid in learning new skills (e.g., St-Clair Thompson & Gathercole, 2006). In practice, when completing an assignment, the ability to retain and apply changes in a teacher’s instructions (working memory), and the ability to inhibit a task-irrelevant impulse (inhibition) are helpful for learning new skills for school-aged children. The aforementioned examples illustrate why EFs have been associated with both math and reading ability in school-aged children and adolescents (see also Ribner, Willoughby, Blair, & Family Life Project Key Investigators, 2017).

Beyond the school-aged years. In a longitudinal study, Moffitt and colleagues (2011) followed a cohort of 1000 children from childhood to age 32. During the first decade of life, several aspects of childhood self-control were combined into a composite from proxy reports, cognitive assessments, and interviews. Self-control is related to EF because it is an aspect of inhibition that is at the level of behavior (i.e., response inhibition). For example, they measured self-control by asking caregivers whether the child “acts before thinking,” has “difficulty waiting [for their] turn,” is “easily distracted,” or has “trouble sticking to a task” etc. Using the composite, they evaluated whether childhood self-control could predict health, wealth, and crime at age 32 after controlling for IQ and social class. They reported that childhood self-control significantly predicted adult health, income, financial planning, and criminal offenses. Moffitt et al.’s (2011) study highlights the importance of EF ability at a young age and its predictive power throughout the lifetime. Possibly related to criminal offenses, EF is associated with risk-taking behavior during adolescence and early adulthood (e.g., Pharo, Sim, Graham, Gross, & Hayne, 2011). Related to job performance, EF has been implicated as being critical for job performance
(Bailey, 2007); and, in participants with ADHD, Barkley and Murphy (2010) reported weaknesses in EF contributed to impairments in job functioning that occur with ADHD.

There are several other aspects of life that may be affected by an individuals’ EF ability. Related to individuals’ social life, EFs have an important role in maintaining social relationships (e.g., Alduncin, Huffman, Feldman, & Loe, 2014; Campbell, Melville, Strutt, & Schall, 2015; Lewis & Carpendale, 2009) and marriages (Eakin et al., 2004). Related to physical health, a recent review by Reinert, Po’e, and Barkin (2013) reported EF has a negative relationship with obesity in children and adolescence (i.e., poorer EF is associated with higher obesity) — extending similar, previously reported research (e.g., Cserjési, Luminet, Poncelet, & Lénárd, 2009; Gunstad et al., 2007). Further, EF is associated with food intake and physical activity in children (Riggs et al., 2010), and deteriorating of gait (e.g., Yoge,v-Seligmann, Hausdorff, & Giladi, 2008). In sum, there are numerous, important outcomes associated with EF which highlight the importance of understanding the profile of EF in DS.

**Measurement of executive function.** Two important factors related to the measurement of EF are the task impurity problem (Burgess, 1997), and the difference between lab-based tasks and proxy reports of EF. Although neither of the issues mentioned above is methodologically relevant to the current paper, they are relevant to the concept of EF and the interpretation of results of the current paper.

**Task impurity problem.** The task impurity problem highlights difficulties in validity when measuring EFs and reliability when measuring EFs. When measuring the components of EF, the validity is suspect because there is no component of EF that can be measured in isolation. Executive functions, by definition, operate in tandem with other cognitive processes; so a substantial amount of variability in different EF tasks is, by definition, related to other cognitive
processes. For example, a rule shift card task requires participants to sort cards based on different characteristics (e.g., color, shape, number of items on a card), and then effectively adapt to a change in sorting rules. However, it also requires the visual perception to identify the cards, verbal ability and WM to understand the directions, the ability to sustain attention throughout the entire task, and other cognitive abilities not directly related to the targeted EF.

Reliability of executive function tasks. The reliability of EF tasks has been called into question in the past for being worse for EF tasks than other cognitive tasks (see Miyake et al., 2000). Consequently, reliability has been investigated in preschoolers (e.g., Beck, Schaefer, Pang, & Carlson, 2011; Gyns & Willis, 1991), school-aged children (e.g., Bishop, Aamodt-Leeper, Creswell, McGurk, & Skuse, 2001), adults (e.g., Lowe & Rabbitt, 1998), and clinical populations (e.g., Tate, Perdices, & Maggiotto, 1998). The results in preschool and school-aged children varied based on the age and task for test-retest reliability (i.e., \( r = .49–.94 \); Beck et al., 2011). However, the majority of test-retest reliabilities are in the acceptable range and higher; there are only a few tasks where reliabilities are low, and it may be specific to the methodology or participants characteristics. More specifically, the test-retest reliability was especially low (≤ .50) for a type of delay task in preschoolers (Beck et al., 2011), and for a planning task in school-aged children (Bishop et al., 2001).

The delay task in children was a “cool” EF delay task used by Beck et al. (2011). It required participants to choose whether the experimenter, or a puppet, would receive a small award immediately, or a larger award after the testing session or the day. For example, “should the puppet get one goldfish now, or four goldfish later?” Their procedure required participants to complete the delay task, do a working memory task, and complete the delay task again with 15min spacing between the two delay tasks. I would argue that their “cool” delay task had low
reliability because it was a poorly designed task for cognitive inhibition in preschoolers. The prepotent response the preschoolers were “suppressing” was the desire for an experimenter or a puppet to have an immediate award. I doubt the validity of the prepotent response in this task—more specifically, I do not think the prepotent response was salient enough that participants needed to actively inhibit it. The study even reported that during this task, most children asked when it would be their turn to get some treats.

Two studies report reliability on the tower of Hanoi task—a task assessing the higher-level EF “planning”—in children. Gnys and Willis (1991) used a 25-minute delay, and their test-retest reliability was $r = .72$; and Bishop et al. (2001) reported a reliability of $r = .50$ after a 30–40 day delay. It is not clear why there is a discrepancy between the two studies; it may be because of the time between tasks and the age of the children—participants were much younger in the Gnys and Willis’ (1991) experiment. In sum, task reliability is adequate or better in all, valid core EF tasks measured in children. The tower of Hanoi, which assesses the higher-level EF planning, had conflicting results related to its reliability in children.

Lab-based tasks versus proxy reports. A recent review conducted by Toplak, West, and Stanovich (2013) evaluated the relationship between lab-based EF tasks and proxy reports. In their review, a lab-based task was defined as a cognitive task administered in person, to the participant, by the experimenter; and a proxy report of EF refers to a questionnaire completed by a parent or teacher. The review paper reported on correlations from 20 studies that had at least one lab-based measure of EF and proxy reports of EF. They reported that only 24% of 286 correlations were statistically significant, and the median correlation was $r = .19$. Further, they noted that these values were likely inflated for a few reasons: (1) not all studies reported nonsignificant correlations, (2) spurious correlations due to a high Type 1 error rate, and (3) the
The file drawer problem might have contributed to many nonsignificant findings never being published. They concluded that lab-based tasks and proxy reports measure different underlying constructs. Further, they made the general conclusion that lab-based tasks measure more processing efficiency of EFs, whereas the proxy reports measure the more goal-oriented aspect of EF. The current meta-analysis only includes lab-based tasks.

**Components of executive function.** When researchers study EF, they often study the components of EF separately. Although it is difficult to “separate” EFs, it is both helpful and theoretically valid to do so. To get a complete picture of EF, I must not only discuss it as a unitary construct but also by its individual components. This paper focuses on the three, well agreed-upon, components of EF in children and adults (i.e., inhibition, shifting, and WM; see Miyake et al., 2000 and Diamond, 2013).

The proposal of the three aforementioned components was supported by latent profile analysis (Miyake et al. 2000) and has subsequently been confirmed with confirmatory factor analysis (CFA; e.g., Friedman, Miyake, Robinson, & Hewitt, 2011). Miyake and Friedman acknowledged that there are other EFs which researchers study; however, these three “have provided useful insights into the nature and organization of individual differences in EFs.” (2012, p. 2). Diamond (2013) supported that these three are generally considered the “core EFs” (p. 2), and clarified that inhibition, shifting, and WM contribute to higher-order EFs (e.g., reasoning, problem-solving, and planning). Diamond’s clarification recognized several other EFs that researchers study and distinguished them from the core EFs. The widespread recognition and usefulness of inhibition, shifting, and WM are the reasons for their inclusion in the current paper.

**Inhibition.** Inhibition is a term that can take on several different meanings, so it is important to operationally define. In the context of EF, inhibition refers to an individuals’ ability
to deliberately suppress a prepotent, or undesired response. For example, if a child is trying to focus while doing their homework, they would demonstrate inhibition if they suppress thoughts and actions unrelated to completing their homework. There is an emphasis on it being a deliberate and controlled process. Simply put, inhibition refers to a persons’ ability to intentionally not respond, act, or attend to something when they do not want to.

An example of a classic inhibition task in children is the day-night Stroop task. First, participants must associate the words “day” and “night” with an unrelated control card (e.g., a purple or green card). After the control trials, they are instructed to say “day” when presented with a black card with a moon on it; and they are instructed to say “night” when presented with a white card with a sun on it. To perform well on the task, participants must inhibit the prepotent response to call the picture with the sun “day” and vice versa.

**Shifting.** Shifting refers to the ability to switch, or shift, between multiple criteria, rules, or tasks as needed. For example, effectively switching from playing a card game with one set of rules to a new a card game with a different set of rules. It is also referred to as “set shifting,” “task switching,” and “attention shift.” Although shifting is only a component of “cognitive flexibility,” many researchers use the terms synonymously (Ionescu, 2012). With regards to EF, shifting is a process that works in tandem with the other components of EF. In essence, shifting involves effectively switching from doing or thinking one thing to another.

An example of a classic shifting task for children is the Dimensional Card Change Sort (DCCS). The DCCS is a simpler version of the WCST. In the DCCS, there are initially two dimensions that cards can be sorted by color and shape. The experimenter demonstrated how to sort the cards by putting them face down in the correct pile (usually based on the color for the first trials) and then instruct the participant on the “sorting rule” for each, individual trial. The
first phase—consisting of six, individual trials—requires participants sort the cards based on one rule. At the end of the first phase, they instruct the participant that the rule changed and they must now sort based on the other dimension. They continue to tell the participant the rule for each trial, but they do not give feedback on whether the participant is correct or incorrect. Some studies have an additional rule change for participants that perform well: they introduce cards with borders and tell them that cards with a border are associated with one rule (e.g., color) and cards without a border are associated with the other sorting rule (e.g., shape).

**Working memory.** Working memory, as conceptualized by Baddeley and Hitch’s model (1974), is a system that maintains, stores, and manipulates information in the mind. For example, children use WM when doing mental math or when they need to remember and apply a series of instructions. Baddeley and Hitch proposed that information could be held in the visuospatial sketchpad (visual) or the phonological loop (auditory). The central executive coordinates the other systems and acts as a supervisory system. Working memory, by their definition, goes beyond a passive storage and retrieval of information. In 2000, Baddeley proposed another component to the widely popular model—the episodic buffer. The episodic buffer was similar in rank to the visuospatial sketchpad and the phonological loop, but it links information across time and modality to form episodes. Therefore, according to the Baddeley model, WM involves the storage, maintenance, integration, and manipulation of multimodal information. In 2012, Baddeley reviewed findings on WM using his framework, and he discussed other conceptualizations of WM as well. He was of the opinion that many of the “competing” theories conceptualized WM very similar to how he had, although other theories differed in terminology and other small nuances.
In the EF literature, there has been a focus on the updating aspect of WM (see Miyake et al., 2000). For example, St Clair-Thompson and Gathercole (2006) treated updating and WM as separate entities but noted they share a common association. It is important to note, many researchers consider Miyake’s et al.’s (2000) inclusion of updating to be analogous to WM (Garon, Bryson, & Smith, 2008).

An example of a classic WM test for children is a dual task. In a verbal dual task, the participant is presented with a list of words (e.g., five words) and is instructed to remember the first word and to tap the table whenever they hear a different word (e.g., “dog”) presented in the list. Another common WM task is the reverse digit span. In the reverse digit span task the experimenter says a list of digits, and the participant is required to repeat the list back to the experimenter in reverse order.

**Other types of executive function.** Although a complete review of all types of EF is outside the scope of this paper, two concepts are relevant to the current meta-analysis. First, there are several higher-level EFs mentioned in the literature (e.g., Diamond, 2013). The higher-level EFs are considered a combination of the three core EFs—inhibition, shifting, and WM—but they are not studied as frequently as the core EFs. Reasoning and problem-solving are both higher-level EFs that are synonymous with fluid intelligence (Diamond, 2016). Planning, however, is a higher-level EF that involves a combination of all the subcomponents to a degree. For example, children and adolescents use planning skills to pack their backpack before going to school. If a child is packing their backpack in the morning, they are using WM to update what they currently have in the bag and what they still need; and they are using shifting to switch between the task of identifying what is needed for the backpack and going to find the items that need to be added.

Second, is short-term memory—especially relevant to individuals with DS. Short-term memory
is not traditionally thought of as a type of EF, but its inclusion in the current project may be
prudent given the poor WM ability in DS (e.g., Jarrold, Purser, & Brock, 2012; Næss, Lyster,
Hulme, & Melby-Lervåg, 2012). Short-term memory (STM) involves the temporary storage and
retrieval of information. Short-term memory is often assessed using recall, or span tasks (e.g.,
digit span, word span, Corsi block-tapping task, etc.).

*Short-term memory.* Short-term memory refers to the temporary storage and retrieval of
information. The duration of STM for the average adult is approximately 15–20 seconds (Revlin,
2012) and the capacity is approximately 5–9 items (Miller, 1956). For example, someone would
use STM to remember a phone number shortly before writing it down. There are strategies that
can be used to aid STM; more specifically, people often use strategies like chunking and
rehearsal. Chunking information involved grouping information into “chunks” or groups of
information to make it easier to remember than an uninterrupted stream of information. For
example, a phone number is easier to remember when broken up into chunks like 205-867-5309
rather than 2058675309. Rehearsal, when used as a STM enhancing strategy, involves the
continual repetition of information that needs to be remembered. For example, the repetition of a
phone number that needs to be remembered until it can be written down.

Related to STM in DS, Bennett, Holmes, and Buckley (2013) conducted a study that used
a computerized STM program for children with DS. They found that the training program
significantly improved performance on trained and non-trained visuo-spatial STM when
compared to a wait-list control group with DS. The improvement was sustained at a four-month
follow-up. This study is important because it highlights that STM is amenable to intervention in
children with DS.
The distinction between short-term and working memory. Short-term and working memory are generally thought of as distinct, and there is evidence to suggest that short-term and working memory operate on different systems (Cowan, 2008). In a landmark study, Engle, Tuholski, Laughlin, and Conway (1999) found that WM—and not STM—showed a strong association with fluid intelligence in college students. This association between WM and fluid intelligence and lack of association for STM and fluid intelligence has been since replicated (Cowan, 2008). However, the picture is clouded because, as Cowan (2008) highlighted in his review paper, there is semantic confusion surrounding the use of the terms “working memory” and “short-term memory.” For example, some consider nonword repetition a WM task and others consider it a STM task. Further, there is experimental data that suggests the concepts are not quite as distinct as previously thought. In 2007, Unsworth and Engle conducted a meta-analytic review on the extent to which WM and STM are distinct constructs when using simple and complex recall tasks. They concluded the extent to which WM and STM are different constructs might depend on the task type and scoring procedures. Additionally, they concluded that simple and complex recall tasks measured the same underlying processes and rejected the notion that WM and STM are different constructs.

In sum, the distinction between WM and STM is clouded by semantic confusion and experimental data. There is a paucity of research on parsing WM and STM early in the development of WM and STM (e.g., in children, or in populations with WM and STM difficulties). For example, does WM ability emerge from STM very early in development? That question cannot be answered with any confidence based on the current data. To be clear, there is not enough evidence to support conclusions that WM and STM are similar or distinct early in
their development. So, as mentioned above, I have included STM as a precautionary component of EF in DS.

**Executive function in Down syndrome.**

Executive function is a known weakness in DS. In fact, it is a weakness in children, adolescents, and adults with DS. Lanfranchi et al. (2010) administered a battery of EF tasks to 15 children with DS, and a TDMA matched group. They reported the DS group performed significantly worse on tasks assessing inhibition, shifting, and WM. They also included tasks assessing fluency, and sustained attention. The sustained attention task showed a similar pattern to the other EF tasks, but performance on the fluency task was not different across groups. Costanzo et al., (2013) conducted a similar study but also included individuals with Williams syndrome. They report similar results—a weakness across the three core components of EF. Additionally, they reported the DS group did not differ from the TDMA group on planning assessed via the Tower of London task.

Rowe, Lavender, and Turk (2006) sought to better understand EF in DS by comparing the performance of individuals with DS to a mental-age matched control group with intellectual disability of unknown etiology. The group with DS reported difficulties in tasks assessing inhibition, shifting, STM, and sustained attention. It is worth noting that Rowe et al. (2006) explicitly stated they did not include tasks of WM because they did not identify any suitable tasks—they were too complex. Instead, they chose to assess STM because the STM tasks tapped aspects of WM, but not the entire concept. This is consistent with the focus of a review by Baddeley and Jarrold (2007) on WM in DS. In their review of WM ability in DS, Baddeley and Jarrold discussed the Baddeley WM model by focusing on tasks assessing STM ability. Baddeley and Jarrold noted that the majority of the applied research in DS on Baddeley’s WM model has
focused on the phonological loop (i.e., short-term memory) more than other components of the model.

Related to STM in DS, verbal STM is a known weakness (see Jarrold et al., 2012 for a review; see Næss et al., 2011 for a meta-analytic review). Næss and colleagues (2011) conducted a meta-analytic review on language and verbal STM in DS and included studies that looked at groups with DS compared to non-verbal TDMA groups. They included seven studies and reported that children with DS were approximately one standard deviation below the non-verbal TDMA groups. Related to visuo-spatial STM, individuals with DS have a mix of strengths and weaknesses (Yang et al., 2014). Yang and colleagues reviewed the literature on several aspects of visuo-spatial STM and challenged the notion that visuo-spatial ability was an area of relative strength in DS. They concluded that spatial sequential memory was commensurate with general intellectual ability in DS; and spatial WM, closure (the ability to combine partial pieces of information into wholes, and vice versa), and wayfinding were weaknesses in DS. Their conclusions are consistent with the fact that performances on Corsi short-term memory tasks have generally not shown weaknesses in DS when compared to control groups of the same cognitive abilities (Jarrold et al., 2012). In a Corsi short-term memory task, the experimenter taps a series of boxes (e.g., on a 4x4 board) in a sequence, and the participant must tap the boxes in the same order. In sum, visuo-spatial STM is not a relative strength—it is either commensurate with general cognitive abilities or a relative weakness. Further, verbal STM is consistently a weakness when compared to groups matched on general cognitive ability.

There is very little information on the unity and diversity of EF in DS, especially at different ages. To date, there have been no latent profile analyses, or exploratory- or confirmatory-factor analyses to identify the structure of EF in DS. Although the current study did
not utilize those methods to elucidate the unity and diversity of EF in DS, they are important methods to be utilized in future research. The results of the current study, however, contribute to the understanding of the unity and diversity in EF by looking at the relationship of different EFs with each other via moderator analyses. For EF in DS, we do not know the exact nature of EF in DS, but it is well-established that there is a general EF weakness in DS. Most studies report a weakness in DS across every aspect of EF they assess in the study (e.g., Costanzo et al., 2013; Lanfranchi et al., 2010; and Rowe et al. 2008), but do not report specific strengths and weakness within the general weakness.

The field would benefit from establishing an EF profile in individuals with DS. To establish an EF profile in DS, we must address several questions: (a) are EF and its components a consistent weakness across all studies? (b) in DS, what is the developmental trajectory for EF and its components? (c) do any components differ from each other (i.e., does a specific pattern of strengths and weaknesses emerge)? (d) does the matching method influence the profile of strengths and weaknesses? It is important to understand because studies on EF in DS use a variety of group matching methods, and they are assumed to be approximately equal. Also, (e) does task modality influence performance on EF tasks? There is an established weakness in verbal STM ability in DS, but less is known about modality for inhibition, shifting, WM, and general EF.

**Benefits of understanding the EF profile in DS.** Understanding the EF profile in DS has a beneficial impact on individuals with DS, their families, researchers, and the community in general. Although there are several benefits, three are especially notable: a better understanding of the EF profile informs current and future interventions; create a frame-of-reference for families, teachers, practitioners, and researchers; and, specific to DS, it provides insight into the
thoughts and behavior of individuals with DS as the field advocates for greater inclusion and independence of individuals with DS.

**Informing interventions.** Executive function is amenable to intervention, especially in children with poor EF (Diamond & Lee, 2011). Diamond and Lee reviewed interventions that aided EF during childhood and early adolescence. They reported that EFs are amenable to interventions, but it is nuanced. Relevant to DS, Diamond and Lee (2011) highlighted a few characteristics associated with the largest EF improvement in interventions. More specifically, individuals with poor EF generally benefit the most, the largest improvements are seen on difficult EF tasks, different interventions are more and less effective at different ages, training appears to transfer but it is somewhat narrow (i.e., specific WM training does not transfer to inhibition tasks, visuospatial training does not transfer to verbal tasks), and the difficulty of the training must increase with participants’ ability to be effective. Among all the aspects that aid interventions, Diamond and Lee (2011) highlighted that there were two commonalities for successful interventions: they did not require students to sit still for a long time, and effective interventions reduce stress, increase social bonding, are enjoyable, improve self-confidence, and are engaging for students.

An EF profile of specific strengths and weaknesses is needed to guide interventions. A better understanding of the EF profile in DS informs the *type* of interventions—helping identify which EFs to target; the developmental *level* of interventions—to ensure they are effective in avoiding floor and ceiling effects; and it helps inform the *timing* of interventions—at what age different types and levels of interventions are appropriate. Informing interventions is critical because, as previously mentioned, EFs are associated with school readiness, academic performance, the transition to adulthood, job success, and other health outcomes.
Adding to the previously discussed outcomes associated with EF, Diamond (2016) highlighted several reasons why EF interventions early in development are especially important. Relevant to DS, Diamond (2016) pointed out that (a) EF problems at a young age do not necessarily improve to developmental norms as they age, rather, the gap may grow larger over time; (b) early interventions affect the academic and vocational trajectory via the positive or negative feedback loop individuals receive throughout development; and (c), although the prefrontal cortex is not fully mature until the mid-20’s, interventions targeting immature biological substrates may promote EFs at the same, or higher levels of proficiency than interventions started once the neural substrates are more mature (i.e., once participants are older). In plain English, the last point highlights that interventions may be as, or more effective for younger children than older ones.

*Creating a frame of reference.* Much research has been conducted in an attempt to understand the development and profile of EFs in TD individuals. Although most researchers may agree that there are still unanswered questions related to EF in TD individuals, the work thus far has created a frame of reference for parents and professionals to use. In DS, however, there is not yet a solid frame of reference related to EF. It is established that EF is a weakness in DS, but the specific nature is not known, and a frame of reference is useful to parents and professionals alike. The current study helped create a frame of reference of EF ability that aids families, teachers, practitioners, and researchers to help individuals with DS.

*Providing insight.* Combining aspects of the previous two points, an EF profile in DS provides insight into how to best help individuals with DS. It helps parents identify specific EFs that may need intervention. It is especially important for parents to understand their child’s specific EF ability because the early family environment plays a critical role in the development
of EF (e.g., Rhoades, Greenberg, Lanza, & Blair, 2011). Understanding specific strengths and weaknesses in DS will help teachers and practitioners understand why certain accommodations for individuals with DS are more and less effective. For example, it could help teachers and school psychologists create effective strategies to help individuals with DS succeed in a classroom environment. An EF profile will help guide researchers on areas of future investigations (e.g., the developmental trajectory of specific EFs, creating interventions, developing classroom strategies, etc.). Last, an EF profile can provide insight for future employers. A better understanding of strengths and weaknesses will help guide employers on which tasks individuals with DS will excel at, and which tasks they will need more guidance to become proficient.

**Appropriateness of a Meta-Analysis**

The main purpose of the current study is to establish the EF profile in individuals with DS. This purpose is well suited for a meta-analysis—better so than another experimental study, or a narrative review. Another experimental study would not have the number of participants, variety in methodology, and could not reasonably answer questions included in moderator analyses (e.g., if the matching method influences the effect sizes). A narrative review suffers from certain pitfalls: they are heavily influenced by significant results, there are no summary effects, and there is no systematic way to generalize conclusions from a large body of literature. A meta-analysis avoids many of the pitfalls that narrative reviews are subject to. The current paper includes a large number of participants in a population that is difficult to recruit. Further, in addition to getting a “summary effect,” the meta-analyst can examine the consistency of effects and evaluate how methodological differences affect the effect sizes via moderator analyses. Please see the Analysis Plan section for more information about meta-analyses.
Hypotheses

The current paper sought to elucidate the EF profile in DS. Generally speaking, I hypothesized that EF in DS would be a weakness when compared to TDMA matched groups. I expected heterogeneity in effect sizes because of the varying study characteristics in our sample (i.e., different ages, matching methods, and modalities); therefore, I had several hypotheses. More specifically, I had three main hypotheses:

1. The overall weighted mean effect size for general EF would be significantly different from zero—in such a way that the DS group would have poorer EF than the TDMA group.

2. Each of the weighted mean effect sizes for the components of EF (i.e., inhibition, shifting, WM, and STM) would be significantly different from zero—in such a way that the DS group would have poorer EFs than the TDMA group.

3. Task-type would be a significant moderator (i.e., whether the task was assessing inhibition, shifting, WM, or STM). More specifically, WM and STM would have significantly larger effect sizes than the other components of EF. Further, I hypothesized that effect sizes for inhibition and shifting would not be significantly different from each other.

In addition to my main hypotheses, I had three exploratory hypotheses.

4. Age, in the group with DS, would be a statistically significant moderator, in such a way that older individuals with DS would have smaller effect sizes than younger individuals with DS. There is some evidence to suggest that general EF improves from adolescence to adulthood in individuals with DS (Loveall, Conners, Tungate, Hahn, & Osso, 2017);
however, the hypothesis is exploratory because the Loveall et al. (2017) study measured EF using a proxy report, and the improvement was not across all EFs.

5. Task modality would be a significant moderator for tasks assessing WM and STM, such that performance on visual modality would have significantly smaller effect sizes than the auditory modality. Evidence supports that verbal-STM is a weakness compared to visual-STM (e.g., Baddley and Jarrold, 2007), and I expect the same pattern to extend to WM.

6. Last, I had two, related exploratory analysis related to the match between groups. I hypothesized that a moderator that tests the use of a general intelligence test versus receptive vocabulary would not be a significant moderator. Phillips, Loveall, Channell, and Conners (2014) did not find differences between a non-verbal intelligence measure and a receptive vocabulary measure. But the analysis remains exploratory because they used a non-verbal measure of intelligence. I also entered matching modality (verbal vs. nonverbal) as a moderator because I expected studies using a verbal matching method to result in significantly smaller effect sizes.
2. METHODOLOGY

Data Collection

Electronic databases. Articles were identified by searches in several electronic databases: PsycINFO, Educational Resources Information Center (ERIC), MEDLINE, Child Development and Adolescent Studies, American Doctoral Dissertations, Proquest, Web of Science, and Academic Search Premier. The following search terms were used in combination with “Down syndrome”: “executive function” or “executive control,” “inhibit*” or “self-control,” “shift*” or “cognitive flexibility,” “visuospatial” or “visuo-spatial” or “visuo spatial,” “phonological,” and “memory.”

Hand search. To ensure a thorough search, I used several “hand search” techniques. Reference lists for the final 57 studies that met inclusion criteria from our database search were backward and forward searched. Additionally, review articles that were included in our search were backward and forward searched. A backward search involves finding relevant articles in the text and references of articles from our initial search. A forward search involves seeing which other papers cited the articles included from our initial search (i.e., I used Google Scholar for the forward search). I searched NIH Research Portfolio Online Reporting Tool for unpublished or ongoing studies meeting the criteria. Last, emails were sent out to first authors of the 57 included articles from the database search to ask about any unpublished research related to DS and EF to address the file drawer problem. All of the hand search techniques yielded two additional studies that met the inclusion criteria.
Inclusion Criteria

All the studies had to meet the following criteria:

1. The study included individuals with DS, and the study must also include a typically developing mental age-matched (TDMA) group.
2. The study included at least one lab-based task that assessed inhibition, shifting, WM, or STM. More specifically, the task needed to meet the task inclusion criteria outlined in Table 1.
3. The study had to be published in or after the year 1990.
4. The article must be published in English, and include the information needed to calculate effect sizes (e.g., inferential statistics or means, standard deviation / standard errors, and the sample size for both groups).

Studies were excluded if more than 50% of the DS sample has mosaicism. This was done because individuals with mosaicism are typically higher functioning than those with full trisomy of chromosome 21. Further, mosaicism accounts for approximately 2% of DS diagnoses (Shin, Siffel, & Correa, 2010), so their EF ability is not a good representation of most individuals with DS.

Figure 1 provides a summary of the search process; the initial database search was on January 24th, 2018. The first two steps of Figure 1 depict the identification stage—using electronic databases and the hand search method. After duplicates were removed, titles were screened based on studies that clearly did not meet search criteria, e.g., “Functional RNAi screen targeting cytokine and growth factor receptors reveals oncorequisite role for interleukin-2 gamma receptor in JAK3-mutation-positive leukemia” (Agarwal et al., 2015). Next, abstracts were retrieved and screened for a preliminary check (e.g., study relevance, whether the research
involves humans, etc.). After abstracts were reviewed, the full-text were retrieved and screened for more detailed information related to the inclusion and exclusion criteria (e.g., full-text unavailable, original data, serial publication, etc.).
Figure 1 – Flow Chart for Search Process

Identified number of published studies
\((n = 6304)\)

Records after duplicates were removed
\((n = 3887)\)

Records after title screening
\((n = 884)\)

Records after abstract screening
\((n = 471)\)

Full-text articles assessed eligibility
\((n = 471)\)

Total number of eligible records
\((n = 57)\)

414 records excluded:
- 170 did not have a TDMA match
- 89 did not have a task meeting our criteria
- 41 were not an empirical article
- 30 Studies did not have participants with DS
- 22 studies were from before 1990
- 22 not enough information to calculate effect sizes (these authors were contacted)
- 18 full-text could not be retrieved
- 15 studies were duplicates (e.g., a dissertation which was later published)
- 7 were not in English

Records after abstract screening
\((n = 471)\)
Table 1 – Task Selection Criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Type of Task</th>
<th>Tasks Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inhibition</td>
<td>The task must require the participant to suppress or avoid a prepotent (automatic) response in order to make the task-required response. There must be a score that reflects the successful suppression of the prepotent response.</td>
<td>Response inhibition, go/no-go, delay of gratification</td>
</tr>
<tr>
<td>Shifting</td>
<td>The task must require the participant to switch between task sets, response rules, or multiple criteria to successfully complete the task. There must be a score that reflects successful switching.</td>
<td>Switching, shifting attention</td>
</tr>
<tr>
<td>Working Memory</td>
<td>The task must require participants to actively maintain, manipulate, or update information and use it when prompted by the experimenter. The information may be visual, auditory, or multi-modal in presentation. There must be a score that reflects the successful maintenance, manipulation, or updating of information.</td>
<td>Updating, manipulation, dual</td>
</tr>
<tr>
<td>Short-Term Memory</td>
<td>The task must require participants to store and recall information after a short-delay (&lt;30 seconds). It may be visual, auditory, or multi-modal in presentation. There must be a score that reflects the successful recall of information.</td>
<td>Free Recall</td>
</tr>
</tbody>
</table>

**Coding**

Every study was coded for moderators and study characteristics by two coders so that reliability could be assessed and reported. Each study was assigned a study ID and had the following information: effect size, sample sizes, calculation procedures (e.g., if means and standard deviations were not reported, effect sizes were transformed from other statistical information), tasks included, and moderator variables. The moderator variables are listed below,
and they are categorized by characteristics associated with the sample, the measurement of EF, and the data source.

**Moderators.**

**Sample characteristics.**

*Age.* I was interested in whether or not there is an association between EF and age in DS. For example, EF decline may be an early indicator of dementia and Alzheimer’s in DS (Fonseca et al., 2016), or individuals with DS may develop strategies to improve EF as they age in comparison to the mental-age matched group. The goal of inclusion of age as a moderator was to add to the knowledge on the trajectory of EF as individuals DS get older.

*Proportion of females with Down syndrome.* I recorded the proportion of participants in the DS groups that are female. There are two primary reasons for including this. First, it helped elucidate whether or not sex has any impact on the strength of effects—which is not expected. Second, it gave us a better understanding of diversity in our sample and how representative they are of the population of individuals with DS.

*Difference in proportion of females.* I noted any difference in the proportion of females between the group with DS and the TDMA group.

*Minority.* I recorded the proportion of participants in the DS groups that are an ethnic or racial minority. There were two primary reasons for including this. First, it helped us understand whether or not being an ethnic or racial minority has any impact on the strength of effects. Second, it gave us a better understanding of diversity in our sample and how representative they are of the population of individuals with DS.
Measurement characteristics.

Executive function task type. I entered the type of EF task (e.g., inhibition, shifting, WM, or STM) as a categorical moderator. Many of the studies investigating multiple EFs report weakness across several EFs (e.g., Costanzo et al., 2013; Lanfranchi et al., 2010; and Rowe et al. 2008), but did not report a consistent pattern of strengths and weaknesses (e.g., if inhibition is a relative strength, and shifting and working memory are weaknesses). Coding task type helped determine the profile of EF in DS because it allowed comparisons of the task types.

Executive function modality. Executive function tasks are presented via the visual modality, auditory modality, or a combination of both modalities. I entered modality as a moderator to evaluate if there are differences across each of the sub-processes of EF and general EF. There is already evidence of a verbal weakness in STM (see Baddeley & Jarrold, 2007 for a review; Lanfranchi, Baddeley, Gathercole, & Vianello, 2012), but other components of EF in DS are less clear.

Matching method. Phillips et al. (2014) suggested there was no difference between the Leiter-R and the Picture Peabody Vocabulary Test-4 (PPVT-4) for matching TD and DS groups. This moderator is concerned with whether using a measure of IQ versus other measures of mental age (e.g., receptive vocabulary) impacts the effect sizes.

Matching modality. Given the known verbal STM weakness in DS (Baddeley & Jarrold, 2007), I collapsed across matching method and categorize whether the matching method was verbal or non-verbal. I am doing this because there are a variety of different tests used in the literature (e.g., Leiter-R, KBIT, WAIS, PPVT, WISC, etc.) that have varying degrees of speaking during administration, and they are all believed to be approximately equal.
Matching closeness. To evaluate the match further, I coded how closely the groups were matched. I calculated the difference between the means of the mental ages reported. I did this under the assumption that there is variability in how closely the matched groups are in mental age across different studies.

Skewness. When possible, I wanted to understand the impact of possible floor and ceiling effects by evaluating low and high skewness. Following the metric outline by Altman and Bland (1996), I computed a skewness variable by taking the difference between the mean and the lowest score for a low skew, or the highest score for high skew, and dividing that by the standard deviation. Scores lower than two indicate skewness. Although this approach is not perfect, it gave us an insight on the degree to which low and high skew impact EF research in DS.

Source characteristics.

Year of publication. Year of publication is commonly included as a moderator in meta-analyses (Decoster, 2009). Although the year of publication rarely has an impact on the strength of the effect, it is wise to include it to rule it out as a confounding variable.

Publication status and funding status. I coded whether or not studies were published in a peer-reviewed journal, and whether or not studies were grant-funded. It is wise to check them because they are both related to the quality of the study. Although the inclusion / exclusion criteria acted as a filter for study quality, it is important to check to see if it is confounding the results (e.g., published studies have significantly larger effect sizes, or if funded studies have consistently smaller effect sizes).

Coding Rules. While coding studies, several tasks had qualities that necessitated creating more explicit coding rules (e.g., multiple outcomes, creative ways of reporting common tasks,
and/or multiple matching measures). Several coding rules were developed so the coding would remain consistent:

1. One effect size per task. To reduce redundancy, only one effect size was computed for each task. For example, if a nonword repetition task reported the results for each syllable, the syllables were averaged.

2. Errors over reaction time. For tasks that report both reaction time and errors as outcomes, I chose errors as the outcome variable. For example, it is common for Stroop tasks to report both errors and reaction times. I made this decision for two reasons, 1) reaction time may be less reliable in children than adults (Brown et al. 2014), and 2) at a mental age of approximately five, errors seem more clinically relevant than reaction time (e.g., parents are more concerned with doing a task correctly than their reaction time during the task).

3. Tasks were excluded if they were redundant with another task. E.g., a study with a digit span, Corsi span, verbal presentation/verbal recall STM task, verbal presentation/visual recall STM task, etc., only the digit span and Corsi tasks were coded.

4. Choose the closest match. When there were multiple matching measures, an effect size was computed, and the smallest effect size (i.e., the closest matching task) was chosen.

5. The task must be associated predominantly with one EF. It is difficult to categorize EF tasks because of the task impurity problem. To reduce inconsistencies in coding this rule excludes tasks that tap different EFs at different points during the task. An example of a task that taps different EFs is one that assesses STM for the first three trials and WM for the last three trials.
Analysis Plan

**Basics concepts of meta-analysis.** In 1976, Glass introduced the term “meta-analysis” by saying “The term is a bit grand, but it is precise, and apt, and in the spirit of ‘meta-mathematics,’ ‘meta-psychology,’ and ‘meta-evaluation.’ Meta-analysis refers to the analysis of analyses. I use it to refer to the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings.” (p. 3). Simply put, a meta-analysis is an analysis of the effects of several studies on the same topic. The purpose of a meta-analysis is the same today as when it was originally outlined by Glass (1976): to provide methodological rigor to the evaluation and summarization of myriad individual studies.

A meta-analysis combines several effect sizes to gain insight into the relationship between two variables. An effect size is a standardized value that reflects the magnitude of an effect (e.g., the difference between groups, correlation, etc.). Statistical significance—and the *p*-value—gives the probability that the researcher would have obtained the result they did, or a greater effect if the null hypothesis were true. Although *p*-values are informative, they do not give any information on the magnitude of the effect, or the “clinical significance.” Effect sizes are the currency of meta-analyses in part because they provide more meaningful information than a *p*-value; an effect size does not answer *whether* there is a difference between groups, but *how much* of a difference there is between groups. This distinction is important because *p*-values are influenced by sample size whereas effect size is not. With a large enough sample, it is possible to find statistical significance with little-to-no clinical significance (i.e., a very small effect size). Effect sizes are almost always more informative than *p*-values are when combined across several studies. Inferential testing using *p*-values is still useful, but it does not provide as much meaningful information as effect sizes do when combining several studies.
**Software.** I conducted the main analyses in Excel, and moderator analysis in R using the package Metafor (Viechtbauer, 2010).

**Computing effect sizes.** When conducting a meta-analysis, the meta-analyst combines the effect sizes from several studies on the same topic and gets a summary effect. For each EF task in the included studies, I computed a standardized effect size (Cohen’s $d$) for the difference between the DS group and the TDMA group. Next, I assigned a study weight to each individual effect size. The study weight is an index of how much a study contributes to the summary effect. The study weight is computed as the reciprocal of the within-study variance for that effect size ($W_i = 1/V_i$). The study weight is, naturally, influenced by standard error; such that, studies with smaller standard errors have greater weight. It is also the case that studies with a larger sample are often more precise—so studies with a larger sample often have a larger study weight (Borenstein, Hedges, Higgins, & Rothstein, 2009).

Additionally, I categorized the tasks into the components of EF (i.e., inhibition, shifting, WM, and STM; see Table 1) and computed summary effects and 95% confidence intervals for each of the EF components. Confidence intervals on a summary effect, or composite effect size, can be interpreted the same way that traditional confidence intervals are—if repeated samples are collected from the population in the same way, the confidence interval contains the true population parameter 95% of the time. Last, I computed an overall summary effect and 95% confidence interval that represent general EF. The summary effect can be thought of as an overall, aggregate effect size. Last, each summary effect is associated with a *p*-value testing whether the effect size is significantly different from zero. I determined if each summary effect is statistically different from zero. This procedure is similar to a highly cited study by Willcutt et al., (2005), and more analogous to a recent meta-analytic review by Snyder (2013).
**Fixed-effect versus random-effects model.** There are several ways a meta-analysis can be performed, and it is largely determined by the topic and studies included in the paper. An important distinction is whether the underlying effect fit the fixed- or random-effect(s) model. The fixed-effect model assumes there is one true effect size underlying all studies in the analyses. It assumes all the factors that could influence the effects in each study are the same (i.e., they are *fixed*), so only measurement error affects the effect sizes. If a researcher is conducting a meta-analysis using the fixed-effect model, then they assume each study act as an individual estimate of the summary effect for the population—the only study-to-study deviations are caused by measurement error. The random-effects model assumes that the effect could vary from study to study due to study-wide participant or methodological differences. When using a random-effects analysis, the meta-analyst assumes different effects sizes in each of the studies that distribute around a mean. So, the effect sizes included in the analysis represent a *random* sample from the population of effect sizes. Importantly, because the random-effects model assumes different effects in each of the studies, the meta-analyst can evaluate whether variables systemically moderate the strength of the effect sizes (e.g., if older participants elicit larger effects).

The analysis in the current paper used the random-effects model because I assumed the studies on EF in DS had different underlying effects (due to sources other than sampling error). More specifically, I assumed the studies in the current meta-analysis had varying effect sizes due to the aforementioned moderators.

**Heterogeneity.** Borenstein et al. (2009) highlight that the point of a meta-analysis is not just to compute a summary effect. Rather, it is to make sense of the pattern of effects. Adapting an example from Borenstein et al. (2009): an intervention that consistently improves receptive
vocabulary by 30% is very different from one that improves receptive vocabulary by 30% on average, with improvement ranging from 0–60%. Heterogeneity in effect sizes is of great interest because it refers to variation in effect sizes. It occurs both because of true variation and random error, and a meta-analysis must parse the true variation from the random to begin to make sense of the pattern of effects. It is important to clarify, when I mention heterogeneity in effect sizes, I am referring only to variation in true effect sizes—not the variation due to random error.

There are several statistics that are important for understanding heterogeneity in a meta-analysis; I will discuss two that are commonly used in meta-analyses. The Q statistic (Cochran, 1954) is the traditional statistic used to assess heterogeneity. The Q statistic uses a p-value to assess the viability that there is no heterogeneity; however, it does not measure the magnitude of heterogeneity. I², on the other hand, compliments the Q statistic because it assesses the proportion of observed variance that is true (not due to error). Scores on I² range from 0–100%, and a score near 0% indicates that nearly all of the observed variance is spurious. I² is useful because it can be compared across different meta-analyses. A drawback to both approaches is that they have to be interpreted with caution when using a small number of studies (k < 20; Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006).

In the current paper, I estimated heterogeneity of effect sizes using the Q statistic. A significant result for the Q statistic indicates that the effect sizes do not represent a single population value, but multiple values (i.e., that the data is better suited for the random-effects model). Further, I reported I², as an index of the magnitude of heterogeneity of the studies included.
**Moderator analyses.** In the random-effects model, the heterogeneity of effect sizes will be evaluated. Heterogeneity is often expected in the random-effects model because it indicates there is more variability in effect sizes than you would expect by chance. If heterogeneity is confirmed, moderator analyses—or a meta-regression—will be conducted to determine if any study characteristics are related to the effect sizes. Moderators can be continuous or categorical and can be entered individually or in combinations. When entered in combination it is similar to using multiple regression. However, instead of assessing the relationship between covariates and a dependent variable (as in multiple regression), the meta-analyst is interested in study characteristics—moderators—relationship with the effect sizes.

For moderator analyses, I entered the moderator variables using the regression-based method outlined by Card (2015) and Borenstein et al. (2009).

**Publication bias.** Publication bias, or the file-drawer problem, refers to the idea that studies that find null, or unexpected, effects are less likely to be published than significant effects. It is also the case that studies that reject the null in the expected direction are more likely to be published than studies that do not. Card (2015) points out that null effects are more likely to be found in “gray literature.” For example, in conference presentations, technical reports, unpublished theses and dissertations, obscure journals, etc.

Publication bias can be managed in several ways. With adequate power, a meta-analyst can code publication status as a moderator and test to see if publication status impacts the strength of the effects. If it is significant, it can then be analyzed separately or corrected. Another common way of evaluating publication bias is by using funnel plots. In a meta-analysis, a funnel plot is a scatterplot of effect sizes relative to their sample size (Card, 2015). Theoretically, there is an approximate line on the funnel plot that an effect size would need to exceed to be
significant—based of the magnitude of the effect size and the same size (for a rudimentary example, see Figure 2). In Figure 2, the dashed line represents the sample size and effect size combination needed for significant results. If publication bias is not present, the effect sizes form an even funnel on either side of the dashed line. If publication bias is present, it would cause an asymmetric funnel plot such that there are notably fewer studies where nonsignificant effect sizes would be expected. Although funnel plots are not perfect, they are frequently used in combination with guidelines (e.g., Sterne et al., 2011) to evaluate publication bias in meta-analyses.

Figure 2 – Example Funnel Plot

There are other approaches worth noting—regression analyses, weighted selection, and using a fail-safe N—but I will only briefly discuss the use of a fail-safe N. A fail-safe N is computed after conducting a meta-analysis. Conceptually, a fail-safe N is the number of theoretically excluded studies (with average effect sizes equal to zero) that would need to be
included to determine that no effect exists. If the number is substantially large, the meta-analyst would feel more comfortable concluding they did not fail to include many unpublished studies, and the conclusions of the meta-analysis are not unduly swayed by publication bias.

For the current meta-analysis, I assessed publication bias several ways. First, I checked to see if publication status was a significant moderator. Second, I checked the funnel plots associated with general EF. Last, I computed a failsafe N for the two overall composites.

**Reliability of coding.** One of the many benefits of a meta-analysis is that it has elements of a review, accompanied with the rigor associated with an experimental study. One of the ways a meta-analysis is more rigorous than traditional review articles are the methods used for findings, selecting, and coding studies. More specifically, it is common for meta-analysis to report the reliability of the coded study characteristics. When coding study characteristics variables, reliability refers to the consistency in the coding scheme. Generally, reliability above .80 for coding study characteristics is acceptable, and any study characteristics with reliability below .70 would need to be explained (Decoster, 2009). Inter-coder reliability was calculated using the agreement rate (number of items agreed on/total number of items).

Card (2015) recommended separating coding items into low- and high-inference coding items. Low inference items refer to items the meta-analyst does not anticipate intercoder coder disagreement (e.g., chronological age). High inference coding refers to items the meta-analyst anticipates possible intercoder disagreement (e.g., EF task type). In the current meta-analysis, there were two coders testing the coding scheme on a 14 studies with 45 effect sizes total. First, they coded 44 out of 45 tasks consistently, meaning there was only 1 task the second coder did not identify as EF tasks. The intercoder reliability for low-inference items (27 items) $\alpha = .95$, and
for high-inference items (8 items), $\alpha = .92$. The intercoder reliability was $\alpha = .94$ for both low- and high-inference items indicating a consistent coding scheme.
3. RESULTS

Across the 57 studies, there were 3,064 participants—1447 in the group with DS and 1617 in the TDMA group. The mean chronological age in the group with DS was 14.76 (SD = 3.34; range = 6.62–40.90), and the mean chronological age for the TDMA group was 5.39 (SD = .93; range =3.27–8.60). Forty-two out of 57 studies reported age-equivalence scores for their matching measure; of those 42 studies, the mean mental age for the group with DS was 5.24 (SD = 1.22), and the mean mental age for the TDMA group was 5.41 (SD = 1.15). Forty-one studies reported the proportion of females with DS and the mean proportion of females was 47.7%; 38 studies reported the proportion of females in the TDMA group which was 47.2%. Only 8 studies reported on the race and ethnicity of their participants and 14% of participants in those studies were non-white minorities. When interpreting mean weighted effect sizes, I used Cohen’s (1988) convention for interpreting small ($d = .2$), medium ($d = .5$) and large ($d = .8$) effects. For the overall composite, I created a composite with and without STM. I created an overall composite without STM to maximize the utility the meta-analysis by separating the traditional, core EFs and the addition of STM. When reporting results the symbol $k$ refers to the number of effect sizes, not the number of studies. Last, effect sizes are coded so that negative effects reflect poorer performance for the group with DS and positive scores reflect better performance for the group with DS.

Weighted Mean Effect Size. To test the first two hypotheses about summary effect sizes being significantly lower than zero, several summary effects were computed. Overall,
participants with DS had significantly lower mean effect sizes for all EF composites—(see Figure 3). All effect sizes were significantly lower than zero ($p < .001$). The overall composite effect size was large for all tasks, $d = -0.74$ ($k = 146$). The overall composite without STM had a medium to large weighted mean effect size, $d = -0.65$ ($k = 68$). For tasks assessing inhibition, there was a medium effect size, $d = -0.51$ ($k = 24$). For tasks assessing shifting, there was a large effect size, $d = -0.80$ ($k = 14$). For tasks assessing WM there was a large effect size, $d = -0.73$ ($k = 30$). Last, there was a large effect size for tasks assessing STM, $d = -0.83$ ($k = 78$).

**Figure 3 – Composite EF Forrest Plot**

![Composite EF Forrest Plot](image)

**Weighted Mean Cohen's $d$ Summary Effect Sizes**

<table>
<thead>
<tr>
<th>Task</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inhibition</td>
<td>$d = -0.51$ [-0.81, -0.22]</td>
</tr>
<tr>
<td>Shifting</td>
<td>$d = -0.80$ [-1.09, -0.50]</td>
</tr>
<tr>
<td>Working Memory</td>
<td>$d = -0.73$ [-0.97, -0.49]</td>
</tr>
<tr>
<td>Short Term Memory</td>
<td>$d = -0.83$ [-1.03, -0.63]</td>
</tr>
<tr>
<td>Overall</td>
<td>$d = -0.74$ [-0.88, -0.61]</td>
</tr>
<tr>
<td>Overall (no STM)</td>
<td>$d = -0.65$ [-0.82, -0.49]</td>
</tr>
</tbody>
</table>

**Heterogeneity.** As expected, there was a large amount of heterogeneity among effect sizes for the overall composite, $Q = 763.51$, $p < .001$; $I^2 = 83.75%$; the $I^2$ value indicates that 83.75% of the variance between studies was due to heterogeneity and not chance. There was also
a large amount of heterogeneity among effect sizes for the overall composite without STM, $Q = 262.02, p < .001; I^2 = 77.85\%$.

**Subgroup Comparisons.** To test the third hypotheses about the relationship between the different types of EF, I used paired-comparisons. Table 2 includes the results for the subgroup comparisons. Only the composite scores inhibition and STM were significantly different from each other ($p = .04$), such that the effect size between participants with DS and the TDMA group was smaller for tasks assessing inhibition than for STM.

**Table 2**

<table>
<thead>
<tr>
<th>Executive Function Subgroup Comparisons</th>
<th>Difference</th>
<th>SE Diff</th>
<th>Z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inhibition vs Shifting</td>
<td>0.28</td>
<td>0.21</td>
<td>1.33</td>
<td>0.18</td>
</tr>
<tr>
<td>Inhibition vs WM</td>
<td>0.22</td>
<td>0.19</td>
<td>1.14</td>
<td>0.13</td>
</tr>
<tr>
<td>Inhibition vs STM</td>
<td>0.32</td>
<td>0.18</td>
<td>1.74</td>
<td>0.04</td>
</tr>
<tr>
<td>Shifting vs WM</td>
<td>-0.06</td>
<td>0.19</td>
<td>-0.32</td>
<td>0.75</td>
</tr>
<tr>
<td>Shifting vs STM</td>
<td>0.04</td>
<td>0.18</td>
<td>0.19</td>
<td>0.42</td>
</tr>
<tr>
<td>WM vs STM</td>
<td>0.10</td>
<td>0.16</td>
<td>0.61</td>
<td>0.54</td>
</tr>
</tbody>
</table>

**Exploratory Analyses.** Table 3 has results for hypotheses four, five, and six. For hypothesis four, age was not a significant moderator ($p = .65$). Regarding hypothesis five, EF task modality was a significant moderator for the overall composite, such that the effect size was substantially larger for verbal EF tasks (see Table 3). The weighted mean effect size for all verbal tasks was $d = -1.01 (k = 91; p < .001)$, and the weighted mean effect size for non-verbal tasks is $d = -.40 (k = 55; p < .001)$. For the overall composite without STM, the mean weighted verbal effect size was $d = -.79 (k = 45; p < .001)$, and the mean weighted nonverbal effect size was $d = -.45 (k = 23; p < .001)$. Also related to hypotheses 5 and 6, matching modality, proximity, and type (IQ vs. Receptive Vocabulary) were not significant moderators. As expected, sex was not a significant moderator either.
Table 3 - Moderator Analyses for Overall Composites

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Moderator</th>
<th>Estimate</th>
<th>SE</th>
<th>Z</th>
<th>p</th>
<th>Lower</th>
<th>Upper</th>
<th>k</th>
<th>R^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Composite</td>
<td>EF Task Modality</td>
<td>intercept</td>
<td>-0.40</td>
<td>0.10</td>
<td>-4.05</td>
<td>&lt;.001</td>
<td>-0.59</td>
<td>-0.20</td>
<td>146</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>-0.61</td>
<td>0.12</td>
<td>-4.89</td>
<td>&lt;.001</td>
<td>-0.85</td>
<td>-0.37</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Match Modality</td>
<td>intercept</td>
<td>-0.86</td>
<td>0.13</td>
<td>-6.57</td>
<td>&lt;.001</td>
<td>-1.12</td>
<td>-0.60</td>
<td>146</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>0.17</td>
<td>0.15</td>
<td>1.09</td>
<td>0.28</td>
<td>-0.13</td>
<td>0.47</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age of DS Group</td>
<td>intercept</td>
<td>-0.64</td>
<td>0.25</td>
<td>-2.62</td>
<td>0.009</td>
<td>-1.13</td>
<td>-0.16</td>
<td>143</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>-0.01</td>
<td>0.02</td>
<td>-0.46</td>
<td>0.65</td>
<td>-0.04</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IQ vs PPVT &amp; BPVS</td>
<td>intercept</td>
<td>-0.78</td>
<td>0.10</td>
<td>-7.55</td>
<td>&lt;.001</td>
<td>-0.99</td>
<td>-0.58</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>0.13</td>
<td>0.16</td>
<td>0.81</td>
<td>0.42</td>
<td>-0.18</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>Overall Composite</td>
<td>EF Task Modality</td>
<td>intercept</td>
<td>-0.45</td>
<td>0.13</td>
<td>-3.37</td>
<td>&lt;.001</td>
<td>-0.71</td>
<td>-0.19</td>
<td>68</td>
</tr>
<tr>
<td>(no STM)</td>
<td></td>
<td>slope</td>
<td>-0.34</td>
<td>0.16</td>
<td>-2.07</td>
<td>0.04</td>
<td>-0.66</td>
<td>-0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Match Modality</td>
<td>intercept</td>
<td>-0.75</td>
<td>0.15</td>
<td>-4.92</td>
<td>&lt;.001</td>
<td>-1.05</td>
<td>-0.45</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>0.14</td>
<td>0.18</td>
<td>0.80</td>
<td>0.43</td>
<td>-0.21</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age of DS Group</td>
<td>intercept</td>
<td>-0.58</td>
<td>0.30</td>
<td>-1.93</td>
<td>0.05</td>
<td>-1.16</td>
<td>0.01</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>-0.01</td>
<td>0.02</td>
<td>-0.31</td>
<td>0.76</td>
<td>-0.05</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IQ vs PPVT &amp; BPVS</td>
<td>intercept</td>
<td>-0.70</td>
<td>0.14</td>
<td>-5.14</td>
<td>&lt;.001</td>
<td>-0.96</td>
<td>-0.43</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>0.14</td>
<td>0.19</td>
<td>0.74</td>
<td>0.46</td>
<td>-0.24</td>
<td>0.52</td>
<td></td>
</tr>
</tbody>
</table>
**Subcomponent Analyses.** Table 4 includes results for moderator analyses on the subcomponents of EF (i.e., inhibition, shifting, WM, and STM). For tasks assessing inhibition and shifting, none of the moderator analyses were significant. However, as hypothesized, EF task modality influenced both WM and STM. For WM, EF task modality was marginally significant \((p = .08)\), so it should be interpreted with caution. The weighted mean effect size for the verbal WM tasks was \(d = -0.92\) \((k = 17; \ p < .001)\), and the weighted mean effect size for nonverbal WM tasks was \(d = -0.54\) \((k = 13; \ p < .001)\). For STM, the EF task modality was statistically significant \((p < .001)\). The mean weighted effect size for verbal STM tasks was \(d = -1.28\) \((k = 46; \ p < .001)\), and the mean weighted effect size for nonverbal STM is \(d = -0.35\) \((k = 32; \ p = .01)\). Last, task modality was a significant moderator for the overall composite that does not include STM \((p = .04)\). Interestingly, the modality of the matching task was a significant moderator in tasks assessing WM; such that, if participants were matched on a verbal task, the mean weighted effect size was smaller. However, it is important to note that there was only one WM task that was matched on nonverbal IQ (i.e., 29 effects sizes with a verbal match and 1 with a nonverbal match). Before interpreting the WM matching modality task as a stable effect, more studies will be needed with a non-verbal matching task.

**Study Quality Factors.** To get an idea of whether there were floor or ceiling effects on EF tasks when studying DS, I created a variable for skewness. I used the technique outlined by Altman and Bland (1996): compute the difference between the mean and the floor (or ceiling) and divide it by the SD. Altman and Bland state that scores lower than 2 indicate skewness. To compute the skewness measure, a task’s minimum and maximum score are needed. Unfortunately, not all tasks are reflected in the skewness measure because many articles did not report the minimum and maximum scores for the tasks, or it could not be deduced from the task.
Table 4 - Moderator Analyses for EF Subscales

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Moderator</th>
<th>Estimate</th>
<th>SE</th>
<th>Z</th>
<th>p</th>
<th>95% Confidence Interval</th>
<th>k*</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inhibition Composite</td>
<td>EF Task Modality</td>
<td>intercept</td>
<td>-0.32</td>
<td>0.27</td>
<td>-1.18</td>
<td>-0.85 - 0.21</td>
<td>24</td>
<td>3.18%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>-0.31</td>
<td>0.32</td>
<td>-0.97</td>
<td>-0.93 - 0.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Match Modality</td>
<td>intercept</td>
<td>-0.59</td>
<td>0.23</td>
<td>-2.60</td>
<td>-1.03 - 0.15</td>
<td>22</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>0.15</td>
<td>0.31</td>
<td>0.47</td>
<td>-0.46 - 0.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age of DS Group</td>
<td>intercept</td>
<td>-0.81</td>
<td>0.60</td>
<td>-1.35</td>
<td>-1.98 - 0.37</td>
<td>23</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>0.02</td>
<td>0.05</td>
<td>0.47</td>
<td>-0.07 - 0.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IQ vs PPVT &amp; BPVS</td>
<td>intercept</td>
<td>-0.57</td>
<td>0.26</td>
<td>-2.20</td>
<td>-1.07 - 0.06</td>
<td>20</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>0.14</td>
<td>0.37</td>
<td>0.38</td>
<td>-0.58 - 0.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shifting Composite</td>
<td>EF Task Modality</td>
<td>intercept</td>
<td>-0.53</td>
<td>0.32</td>
<td>-1.67</td>
<td>-1.15 - 0.09</td>
<td>14</td>
<td>1.44%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>-0.36</td>
<td>0.36</td>
<td>-1.00</td>
<td>-1.06 - 0.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Match Modality</td>
<td>intercept</td>
<td>-0.95</td>
<td>0.20</td>
<td>-4.82</td>
<td>&lt;.001 - 0.56</td>
<td>14</td>
<td>10.89%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>0.38</td>
<td>0.29</td>
<td>1.32</td>
<td>0.19 - 0.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age of DS Group</td>
<td>intercept</td>
<td>-0.24</td>
<td>0.64</td>
<td>-0.38</td>
<td>0.71 - 0.19</td>
<td>13</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>-0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.34 - 0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IQ vs PPVT &amp; BPVS</td>
<td>intercept</td>
<td>-0.71</td>
<td>0.35</td>
<td>-2.03</td>
<td>-1.40 - 0.03</td>
<td>12</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>-0.06</td>
<td>0.40</td>
<td>-0.14</td>
<td>0.89 - 0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Working Memory Composite</td>
<td>EF Task Modality</td>
<td>intercept</td>
<td>-0.53</td>
<td>0.16</td>
<td>-3.27</td>
<td>0.001 - 0.85</td>
<td>30</td>
<td>18.48%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>-0.39</td>
<td>0.23</td>
<td>-1.73</td>
<td>0.08 - 0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Match Modality</td>
<td>intercept</td>
<td>-2.42</td>
<td>0.65</td>
<td>-3.71</td>
<td>&lt;.001 - 3.71</td>
<td>30</td>
<td>23.14%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>1.71</td>
<td>0.66</td>
<td>2.58</td>
<td>0.01 - 0.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age of DS Group</td>
<td>intercept</td>
<td>-0.57</td>
<td>0.38</td>
<td>-1.51</td>
<td>0.13 - 1.33</td>
<td>30</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>-0.01</td>
<td>0.02</td>
<td>0.45</td>
<td>0.65 - 0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IQ vs PPVT &amp; BPVS</td>
<td>intercept</td>
<td>-0.79</td>
<td>0.18</td>
<td>-4.40</td>
<td>&lt;.001 - 1.15</td>
<td>22</td>
<td>2.77%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>0.39</td>
<td>0.33</td>
<td>1.21</td>
<td>0.23 - 0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short-Term Memory</td>
<td>EF Task Modality</td>
<td>intercept</td>
<td>-0.35</td>
<td>0.13</td>
<td>-2.70</td>
<td>0.007 - 0.61</td>
<td>78</td>
<td>36.55%</td>
</tr>
<tr>
<td>Composite</td>
<td></td>
<td>slope</td>
<td>-0.90</td>
<td>0.17</td>
<td>-5.23</td>
<td>&lt;.001 - 1.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Match Modality</td>
<td>intercept</td>
<td>-0.99</td>
<td>0.20</td>
<td>-4.85</td>
<td>&lt;.001 - 1.39</td>
<td>78</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>0.22</td>
<td>0.24</td>
<td>0.91</td>
<td>0.36 - 0.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age of DS Group</td>
<td>intercept</td>
<td>-0.89</td>
<td>0.38</td>
<td>-2.31</td>
<td>0.02 - 1.65</td>
<td>77</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>0.00</td>
<td>0.02</td>
<td>0.16</td>
<td>0.87 - 0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IQ vs PPVT &amp; BPVS</td>
<td>intercept</td>
<td>-0.87</td>
<td>0.14</td>
<td>-6.04</td>
<td>&lt;.001 - 1.15</td>
<td>66</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slope</td>
<td>0.11</td>
<td>0.24</td>
<td>0.45</td>
<td>0.65 - 0.36</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
description. Table 5 includes information for the skewness measure: the means and standard deviations, the percentage of effect sizes that are below the threshold of 2, and the number of tasks included in the computation of the skewness measure.

Table 5 shows that for all EF tasks, 40% of tasks indicated the group with DS was skewed towards the floor of the task, and 53% of tasks indicated the TDMA group was skewed towards the ceiling of the task. Excluding tasks assessing STM, 57% of tasks in the group with DS was skewed towards the floor and 71% of tasks the TDMA group was skewed toward the ceiling. Tasks assessing shifting and WM may have been especially susceptible to low skew effects in the group with DS (73% and 62% below 2), and high skew effects in the TDMA group (100% and 64% of scores below 2). The combination of low skew effects in DS and high skew effects in the TDMA group indicates that effect sizes may be systematically smaller for shifting and WM tasks. To better understand the differences between groups, two t-tests were conducted using effect sizes for all EF tasks. The first, Welch’s t-test for unequal variance, compared DS low skew to TD low skew and found a significant difference \( t(153.33) = -3.23, p = .001 \); the second, student’s t-test, compared DS high skew to TD high skew but did not find a significant difference \( t(84) = -.67, p = .50 \).

Match proximity (i.e., how well the groups were matched), was entered as a moderator variable for all composite scores. It was marginal for tasks assessing shifting \( (p = .08) \) such that the closer the groups were matched the smaller the effect sizes were (the group with DS had a lower mean MA on average for shifting tasks). However, because there were only 14 studies it is likely the p-value is unreliable, and it is also only marginally significant. Further, match proximity was not a significant moderator for any other composite score.
Table 5 - Mean Low and High Skew, and Percentage of Skewed Scores

<table>
<thead>
<tr>
<th></th>
<th>Floor N</th>
<th>Ceiling N</th>
<th>DS Low</th>
<th>%</th>
<th>TD High</th>
<th>%</th>
<th>DS High</th>
<th>%</th>
<th>TD Low</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>128</td>
<td>49</td>
<td>2.64 (1.42)</td>
<td>40%</td>
<td>3.15 (4.03)</td>
<td>53%</td>
<td>3.69 (3.46)</td>
<td>27%</td>
<td>3.84 (3.85)</td>
<td>19%</td>
</tr>
<tr>
<td>Inhibition</td>
<td>16</td>
<td>9</td>
<td>2.54 (1.61)</td>
<td>38%</td>
<td>5.83 (7.86)</td>
<td>67%</td>
<td>4.68 (6.41)</td>
<td>67%</td>
<td>2.75 (1.42)</td>
<td>38%</td>
</tr>
<tr>
<td>Shifting</td>
<td>11</td>
<td>5</td>
<td>1.95 (1.23)</td>
<td>73%</td>
<td>1.27 (0.33)</td>
<td>100%</td>
<td>2.93 (1.23)</td>
<td>20%</td>
<td>2.24 (0.94)</td>
<td>55%</td>
</tr>
<tr>
<td>Working Memory</td>
<td>26</td>
<td>14</td>
<td>2.06 (1.18)</td>
<td>62%</td>
<td>2.14 (1.10)</td>
<td>64%</td>
<td>3.49 (2.67)</td>
<td>21%</td>
<td>2.94 (1.53)</td>
<td>31%</td>
</tr>
<tr>
<td>Short-Term Memory</td>
<td>69</td>
<td>15</td>
<td>2.99 (1.39)</td>
<td>30%</td>
<td>3.11 (2.32)</td>
<td>40%</td>
<td>3.54 (2.14)</td>
<td>20%</td>
<td>4.69 (4.82)</td>
<td>6%</td>
</tr>
<tr>
<td>Overall (-STM)</td>
<td>53</td>
<td>28</td>
<td>2.18 (1.33)</td>
<td>57%</td>
<td>3.17 (4.74)</td>
<td>71%</td>
<td>3.77 (4.03)</td>
<td>36%</td>
<td>2.74 (1.40)</td>
<td>38%</td>
</tr>
</tbody>
</table>

Note: for the percentages, the skewness cutoff was 2; scores below 2 indicate skewness.
**Diagnostic Analyses.** I evaluated publication bias using three methods: moderator analyses, funnel plot analyses, and by computing a fail-safe N. For the overall composite, moderator analyses for publication status ($p = .33; k = 146$) and funding status ($p = .55; k = 146$) were both nonsignificant. For the overall composite without STM, publication status ($p = .73; k = 68$) and funding status ($p = .99; k = 68$) were both nonsignificant. I used Egger’s regression (Egger, Smith, Schneider, & Minder, 1997) as an objective measure of funnel plot asymmetry; following recommendations from Zwetsloot et al. (2017) the inverse of the square root of the sample size (i.e., $1/\sqrt{n}$) was used as the sample size-based precision estimate with Cohen’s $d$; this was done to avoid the overestimation of bias introduced when using standard error along with Cohen’s $d$ in funnel plot analyses. For the overall composite, Egger’s regression for funnel plot asymmetry was marginally significant $z = -1.816, p = .06$, indicating that the funnel plot was marginal asymmetrical. For the overall composite without STM there was not significant evidence for funnel plot asymmetry, $z = -1.62, p = .10$. Last, I computed Rosenthal’s (1979) fail-safe N for the overall composite: $N = 54,122, p < .001$, and for the overall composite without STM: $N = 9600, p < .001$. Taken together, the measures of publication bias indicate there may be slight publication bias present, but the publication bias present cannot account for the observed difference between groups in the current meta-analysis.

**Common language effect size.** To put the overall weighted mean effect size of $d = -.74$ into context, it is helpful to think about it a few different ways. Because Cohen’s $d$ is not immediately interpretable to many people, it is helpful to present the overall findings of in terms of overlapping populations and odds. Cohen’s $U_3$ (1988) is a measure of overlap between two populations and a Coden’s $d$ of -.74 means that approximately 77% of the TDMA group was above the mean of the group with DS. Another way to think about the overall summary effect is
by converting Cohen’s $d$ to the common language effect size statistic (Ruscio, 2008). In this meta-analysis, the common language effect size statistic of approximately 70% (converted from $d = -.74$) represents the probability that an individual with DS will have poorer EF than a TDMA match person. That is, there is a 70% chance that a randomly selected person with DS will have poorer EF than a randomly selected typically developing individual matched on mental age.
4. DISCUSSION

The current meta-analysis is the first to establish the profile for EF in DS, and to synthesize the magnitude of the deficit compared to TDMA children. Overall, the meta-analysis revealed that individuals with DS performed significantly poorer on tasks assessing EF when compared to a TDMA matched group. More specifically, the summary effect sizes for tasks assessing inhibition, shifting, WM, STM, and for both overall composites were significantly below zero. Previously, it was understood that EF was a weakness for individuals with DS, but the exact nature of the weakness was unclear. There was conflicting evidence on whether or not shifting and inhibition were relative strengths, or commensurate with DS’s WM difficulties. This meta-analysis solidifies a profile for lab-based tasks of EF in DS for future researchers to clarify and build upon. One of the main contributions is establishing the magnitude of the EF deficit compared to a TDMA group is consistently a medium to large effect size. The current paper also provides some insight into the relationship between the components of EF—namely, inhibition may be a relative strength for individuals with DS. Subgroup comparisons revealed individuals with DS perform significantly better on tasks assessing inhibition than tasks assessing STM, but they did not reveal any other significant differences. In sum, there is a clinically significant overall weakness in EF for individuals DS with a relative strength in inhibition ability.

There was a large degree of heterogeneity in the overall composites (i.e., \( I^2 \approx 80\% \)), indicating that there is a large amount of true between-study variance. Although moderator analyses accounted for some of the heterogeneity, there is still a large portion that is unaccounted
for. There are several possible explanations for the heterogeneity; the first being that a large amount of heterogeneity was expected. Because I included several different cognitive abilities in the same analysis, I expected there to be differences. There were also issues related to the measurement of EF. As outlined in Table 5, the presence of low- and high-skew would introduce noise into the measurement of different EFs. For example, effect sizes from studies that selected tasks that are skewed will likely be different from studies with more sensitive measures. Additionally, EF tasks in DS are often simplified from other EF tasks in another population, or there are several versions of the same task (e.g., the animal Stroop vs. the day-night Stroop). The lack of consistency in measures likely contributed to the heterogeneity. In fact, the task that appeared most frequently in the current meta-analysis were both STM tasks: the digit span task, and the Corsi task. Both the digit span task and the Corsi task are usually administered the same way across different studies. Short-term memory had a larger portion of heterogeneity explained by moderator analyses than the other components of EF. The consistency in administration of STM tasks likely contributed to the larger portion of explained heterogeneity.

Moderator analyses revealed that EF task modality influences the overall composites and the composites for tasks assessing WM and STM. As expected, the results support an overall visual strength, which is driven by WM and STM tasks. Moderator analyses did not reveal a significant effect for age, matching proximity, or matching task (IQ vs. Receptive Vocabulary). Related to study quality factors, exploratory analyses revealed that effect sizes might be slightly understated in the current review because of the presence of skew towards the floor in the group with DS and skew towards the ceiling in the TDMA group. An inspection of the percentage of tasks below the cutoff value for skewness in Table 5 indicates that the pattern is stronger in tasks assessing shifting and working memory. This pattern of skewness indicates that effect sizes may
be underestimated for EF tasks when compared to TDMA groups—especially tasks assessing shifting and WM. Researchers should keep this in mind when interpreting the results of this meta-analysis, and when creating tasks for individuals with DS.

**Implications for EF in DS Theory.** The results of the meta-analysis are mostly consistent with prior literature, which makes sense. But there is one inconsistency with previous literature that I want to highlight—is shifting a strength or weakness? The meta-analysis reported a mean effect size of \( d = -.80 \), indicating a large difference in performance compared to the TDMA reference group. However, Daunhauer et al. (2014) and Loveall et al. (2017) reported that it might be a relative strength in DS. I believe the discrepancy arises because this meta-analysis only included lab-based tasks compared to a TDMA group, and the two aforementioned papers used proxy reports to measure shifting ability. However, other aspects of EF are consistent with previous literature. For example, Godfrey and Lee (2018) conducted a review examining effect sizes across age for STM, WM, and long-term memory in DS. They compared individuals with DS to TDMA match groups and to individuals with intellectual disabilities. They reported an average effect size of \( d = -.81 \) for WM (verbal: \( d = -1.00 \), and visual: \( d = -.58 \)), and an average effect size of \( d = .91 \) for STM (verbal: \( d = -1.40 \), and visual: \( d = -.32 \)). Godfrey and Lee’s findings are very similar to the finding reported in the current paper.

Toplak et al.’s (2013) review highlighted that lab-based tasks and proxy reports likely measure different aspects of EF. Daunhauer and colleagues (2017) conducted a study comparing lab-based tasks and a proxy measure for EF in DS, and they found moderate correlations for some EFs but not all. Daunhauer et al.’s study provided support for Toplak’s conclusions in a population with DS: proxy measures tap more goal-oriented aspect of EF, and that cognitive measures tap more cognitive efficiency aspects of EF. In sum, the difference between lab-based
tasks and proxy measure help explain the discrepancy caused by previous reports of shifting being a relative strength in DS. Further, because the current meta-analysis only includes lab-based tasks, the proposed EF profile represents individuals with DS’s cognitive efficiency related to EF better than goal-oriented behavior related to EF.

Although I originally coded for lab-based tasks, there were only 2-3 studies that met inclusion criteria that used proxy reports. Further, some studies that use proxy reports choose to use a report corresponding with participants’ mental ages, and others use proxy reports corresponding with their chronological ages. Because there were few studies that met inclusions criteria, we just included lab-based tasks to be consistent throughout the paper.

**Informing Interventions.** One of the benefits to understanding the EF profile in DS is informing interventions. Although an intervention targeting any aspect of EF would be beneficial—considering there is a deficit in all EFs measured—it may be especially beneficial to target shifting, verbal WM, and verbal STM. There has been small to moderate success in interventions targeting WM and STM in individuals with DS (e.g., Conners, Rosenquist, Arnett, Moore, & Hume, 2008), but I am not aware of any intervention targeting shifting in individuals with DS. As far as informing the level of intervention, for all EFs it would be prudent to target an age equivalence of ~5-year-old. The mean mental age for our group with DS was 5.26, and EFs in the group with DS were lower than the mental-age matched group, so a target age equivalence of ~5 years old for a typically developing child seems appropriate for most age groups in DS.

Understanding the appropriate mental age for an intervention is critical because the intervention must be appropriately difficult / achievable for the individuals participating in it. For example, if a 5-year-old participated in an intervention targeted for 3-year-old they may not benefit because they are operating well above a 3-year-old level. However, if a 5-year-old
participated in an intervention targeted for 7-year-olds, they may find the material too difficult and, consequently, not benefit from the intervention. Last, the moderator analyses did not elucidate any relationship between age and EF in DS. So, for the timing of interventions, I will adopt Diamond’s (2016) reasoning highlighting the importance of early intervention until there is empirical evidence in DS showing ideal timing for interventions. However, there will need to be more interventions implemented before we can conclude with any confidence that early interventions are advantageous for EF in individuals with DS.

Providing Insight. An understanding of the strengths and weaknesses of the components of EF in DS helps create a frame of reference for what is typical for individuals with DS. For example, if a worker has a new employee with DS, they can expect their new employee will require extra training in tasks that require STM, but not as much training in a task that requires inhibition. A better understanding of the strengths and weaknesses in the EF profile will also aid school psychologists, clinicians, and teachers create more effective interventions and accommodations in the classroom. A school psychologist and behavior interventionist should expect to adapt their classroom interventions that require EFs to a mental age of ~5 years old for teenagers with DS. Teachers should be aware that activities that use STM, WM, or shifting will require extra instruction time for their students with DS. Last, parents and caretakers of individuals with DS can have a better understanding of their child’s abilities, and what types of activities may be more difficult for individuals with DS.

The proposed profile also has implications for social interactions as well. For example, because of the weakness in shifting it may be helpful not to jump conversations topics too many times while talking with individuals with DS. Given the visual strength overall and especially on short-term and working memory tasks, it may be more appropriate to give directions or
demonstrate new concepts with visual aids. Individuals with DS may also benefit reminders while they are completing goal-oriented tasks.

**Limitations.** This meta-analysis is limited in a few ways. First, the moderator analyses for inhibition and shifting may have been nonsignificant because of a lack of power. Marín-Martínez and Sánchez-Meca (1998) found that moderator analyses with \( k < 20 \) may be unreliable. Because there were fewer studies with inhibition and shifting, I cannot definitively say whether or not the moderators impacted the differences between groups. Second, the preference for using errors as a dependent variable over reaction time (when there was the option) could be a limitation depending on the reader’s preference—errors or reaction time. The relative strength in inhibition may disappear if reaction time is used as the outcome for tasks like the Stroop. Although I did not test it, anecdotally, the reactions times seemed to elicit larger effect sizes for EF tasks. I felt theoretically justified in the use of reaction time as an outcome for individuals with a mean mental age of ~5-6, but it may have biased the results. Third, a meta-analysis is an excellent tool to begin to establish the EF profile in DS, but further work is needed to solidify the pattern of strengths and weaknesses in the current paper. There is a definite need for, more studies assessing shifting and inhibition. Last, there are two additional analyses that were planned but not conducted for this analysis because of a lack of power: there were only three studies that used proxy measures of EF and only four effect sizes that used a behavioral assessment of inhibition.

**Future Directions.** When thinking about future directions I think it is important to be driven by questions the field needs to answer. In the introduction I highlighted several questions related to EF in DS: (a) are EF and its components a consistent weakness across all studies? (b) in DS, what is the developmental trajectory for EF and its components? (c) do any components
differ from each other (i.e., does a specific pattern of strengths and weaknesses emerge)? (d) does the matching method influence the profile of strengths and weaknesses? Also, (e) does task modality influence performance on EF tasks?

The results of the current meta-analysis provided information to help answer several important questions related to the field, especially related to EF ability, its profile, and the measurement of EF in DS. The current meta-analysis also highlighted several areas for future research to explore. Namely, more work needs to go into creating and understanding effective interventions, evaluating measures of EF for individuals with DS, and conducting follow studies—especially in inhibition and shifting—to elucidate the profile of EF in DS further.

Future research should explore the consistency of the EF profile proposed in the current study. The proposed profile, by nature, is an aggregate for all studies related to EF that met the inclusion criteria, so it provides very useful information. However, it is not clear how consistent the profile within studies. For example, if there is a study that includes 50 individuals with DS, do 45/50 display this pattern of strengths and weakness in EF, or are there several profiles—e.g., a few participants have great shifting but poor inhibition, or poor shifting but a WM strength—which combine to create the proposed profile from the current meta-analysis. A future study that employs a technique like confirmatory factor analysis could elucidate the consistency of the proposed profile. More work should also be conducted to understand the developmental trajectory of EF. It is clear that the components are distinct: STM and inhibition are significantly different. But at what age do the components start to differ? Additionally, at what age do STM and WM ability start to differ? In typically developing children, there is evidence supporting that EFs become more distinct with age, and there are general ranges for which EFs emerge. Our understanding of EF in individuals with DS would benefit from the same types of clarifications.
As previously mentioned, there have been studies that looked at the effect of WM and STM training in DS, but there are very few studies targeting inhibition and shifting. Diamond and Lee (2011) highlighted that effective interventions do not require sitting for long periods of time, and they reduce stress, increase social bonding, are enjoyable, improve self-confidence, and are engaging. It is critical that researchers, teachers, clinicians, and interventionists work together to develop interventions that improve EF in DS. With the increased push towards the independence of individuals with DS, the processes that regulate goal-oriented behavior need to be bolstered at an early age. Further, more work needs to be conducted to identify at what time interventions that target EF are appropriate for individuals with DS. It is likely that early intervention is key for DS as it is in many other populations (Diamond, 2016), but more work needs to be conducted to confirm the benefit of early intervention in DS.

In the field of intellectual and developmental disabilities, including DS, there is an increased focus on identifying good, sensitive outcome measures. In 2016, Hessel et al. reported preliminary results on using the NIH Toolbox for groups with intellectual disabilities. The NIH Toolbox is a new collection of computerized cognitive assessments—a toolbox, if you will—that is appropriate to use in individuals with intellectual and developmental disabilities. In 2017, Budimirovic et al. reported on appropriate outcome measures for clinical trials in fragile X syndrome. A month earlier in 2017, Esbensen et al. published a paper that started the process of establishing agreed-upon measures for clinical trials in individuals with DS. The current paper further highlights this need because of the evidence of skewness on EF measures in DS. Future research should focus efforts on exploring more sensitive measures for shifting and WM. Sensitive, psychometrically sound measures are the foundation for identifying effective interventions.
**Conclusions.** It is clear that there is an EF weakness in DS, even when compared to children of the same mental age. It is also clear that there is a visual strength in STM and WM and relative strength in inhibition. More research needs to be conducted to completely understand the EF profile in DS, and future work should also prioritize creating and validating suitable measures for EFs in DS, which will aid in assessing the effectiveness of future interventions. Last, understanding the profile of EF will allow other researchers, clinicians, teachers, and parents to better help individuals with DS. An established profile for EF informs the type and level of future interventions, and it also provides insight into the abilities of individuals with DS. Ultimately, a better understanding of the EF profile in DS will lead to improved interventions, it will increase the quality of accommodations and support, and it will improve the quality of life for individuals with DS and their families.
REFERENCES


