UNSTRUCTURED PERFOMANCE TASK TO ASSESS EXECUTIVE FUNCTIONS: A STUDY IN TYPICALLY-DEVELOPING CHILDREN

ELIZABETH ARDEN WANSTALL

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ARTS

GRADUATE PROGRAM IN PSYCHOLOGY
YORK UNIVERSITY
TORONTO, ONTARIO

AUGUST 2019

Abstract

A weak concordance between performance-based measures and behavioural ratings of executive functions (EF) has been well-documented in children with and without neurodevelopmental conditions. Performance-based EF measures are administered under highly structured conditions and may not reflect children's performance in everyday environments where less guidance may be provided. The Unstructured Performance Task (UPT) is a novel performance-based EF task designed to include 42 easy problems that are randomly placed on a large sheet of paper (11" x 17") and are administered with minimal direction and external monitoring. The psychometric properties and correlates of the UPT and the UPT-2 (i.e., an updated version of the UPT) were investigated in this study. The UPT was found to have good psychometric properties and scores were significantly related to children's academic abilities. The UPT-2 was then examined in a subsample of the original community sample of children plus a sample of new participants. The UPT-2 generated greater variability in scores and demonstrated improved psychometric properties. The UPT-2 was also found to be significantly related to performance-based tasks of EF and academic abilities. Overall, results indicate that the UPT/UPT-2 is a promising measure of EF performance in children.

Acknowledgements

First, I would like to sincerely thank my supervisor, Dr. Maggie Toplak. I am very grateful for the unwavering support and guidance that I receive from you. Your endless enthusiasm and dedication to your work constantly inspires and motivates me. To everyone in the Toplak Research Lab – Stella, Wafa, Jala, Josh and my wonderful lab twin Rachael – I feel very fortunate to work alongside such kind and intelligent people, and thank you for all of your insight and help along the way. I am looking forward to continuing my work in this amazing lab!

Secondly, I would like to extend my thanks to the members of my committee. Dr. Rhonda Martinussen – thank you so much for your generosity and willingness to collaborate on this project. Your insight and contributions to this work have been invaluable. It has been such a pleasure to work with everyone at the Learning, Engagement and Attention Lab and the Dr. Eric Jackman Institute for Child Studies, with a very special thank you to Joanna! Dr. Melody Wiseheart and Dr. Caroline Davis – thank you both for your constructive feedback on this thesis.

Thirdly, I am endlessly grateful for the love and support of my family and friends. Thank you to my incredible parents, who have lovingly and fiercely supported all of my dreams and ambitions from the very beginning. Je vous aime énormément. Mer and Nico – I am so lucky to have siblings that I can also call my best friends. I know that I can count on you for anything, and that means the world to me. I am also very fortunate to have such an amazing extended family, with my wonderful grandparents at the center of it all. Grandma and Poppy – thank you for showering me with love and wisdom, and for being my home away from home. I also want to thank my cohort – Ben, Bianca, Khad, Kyla, Laura, Rachael, Sam and Tracy – as well as our adoptees, Holly and Jenkin. You are all unbelievably brilliant and kind, and I feel so privileged to share this journey with you. Even though it hasn't always been easy, it is what it is, and we got through it together! To my music family, affectionately referred to as "Octava and friends" – thank you for making music with me when I most needed it, and for sharing a few laughs along the way. Thank you to "my Maggie" for being my cheerleader and closest confidante – your strength and determination never cease to inspire me. And to my loving partner Kody – thank you for your unfaltering patience and support, even when things get tough. You are my rock.

Finally, I would like to thank all of the children and parents who participated in this study. I sincerely appreciate all of the time and experience that you generously shared with me.

Table of Contents

List of Tables

List of Figures

Unstructured Performance Task to Assess Executive Functions: A Study in Typically-Developing Children

Executive functions (EFs) are critical abilities used for problem-solving and goal-directed behaviour (Diamond, 2013). There is some contention regarding the nature of the definition of EFs, with various theories proposing different models. Generally, the dominant perspective is to view EFs as a multidimensional construct, acknowledging overlapping and separable components of these abilities (Engle, 2018; Miyake et al., 2000). EFs are currently measured using performance-based tasks as well as behavioural rating measures. Despite the obvious intention of both tasks to measure EFs, there is a surprisingly low rate of convergence between these two types of measures (Toplak, West & Stanovich, 2013). It has been posited that this may in part be due to the amount of structure and direction provided to examinees during the administration of performance-based measures of EF. For example, EF tasks are typically administered in highly structured environments in order to assess optimal performance (Toplak et al., 2013). In contrast, behavioural rating scale assessments of EF require observers (e.g., parents/teachers) to rate how well a child is able to engage in various EF tasks during regular activities where there may be little guidance or structure available to support the implementation of the EF (e.g., organize a binder). The Unstructured Performance Task (UPT) of EF was developed to be a performance-based EF task that provides little externally guided structure to the participant completing the task (Ledochowski et al., 2019). To advance this work, the aim of this project was to replicate data patterns and refine the measure itself. The aim of Study 1a was to examine the psychometric characteristics of the UPT from the original pilot data on this task. The aim of Study 1b was to replicate and extend the findings by Ledochowski et al. (2019) in a community sample of children, particularly to guide the development of an updated version of the UPT that addresses some of the limitations of the original version. Finally, the aim of Study 2 was to pilot the updated version of the UPT in a community sample of children.

Executive Functions

Executive function (EF) is an umbrella term used to encompass the cognitive abilities necessary for problem-solving and goal-directed behaviour (Diamond, 2013). Many current definitions explain EF abilities as being high-level cognitive processes that enable successful goal-directed behaviour (Friedman & Miyake, 2017). It is generally accepted across literatures that EFs can be viewed as a multidimensional construct with overlapping as well as separable components to these abilities (Miyake et al., 2000). Diamond (2013) explains that there is general consensus regarding three core EFs: inhibitory control, working memory, and cognitive flexibility. *Inhibitory control* refers to the process of stopping a response and the ability to cancel an ongoing action or thought. This can include self-control (i.e. behavioural inhibition) as well as interference control (i.e. selective attention and cognitive inhibition) (Diamond, 2013; Schachar et al., 2007). *Working memory* is the ability to manipulate and temporarily store information in day-to-day life (Kane et al., 2007). This can involve replacing and updating old information with new and relevant information for a given task being performed (Jewsbury, Bowden & Strauss, 2016). *Cognitive flexibility* is the ability to flexibly change perspectives or approaches to a problem or situation, by seeing other people's perspectives or other temporal spatial perspectives. It also includes adjusting to updated demands, rules, or priorities. In many ways, cognitive flexibility requires and builds on both inhibitory control and working memory. For example, to be cognitively flexible to change one's perspective, this requires the inhibition of one's previous perspective and the activation of working memory to load a new perspective (Diamond, 2013). Cognitive flexibility is closely related to set-shifting (i.e. the ability to cognitively switch tasks, use alternative strategies, and process information from various sources) (Zelazo, Craik & Booth, 2004).

However, varying theories and conceptualizations of EF have been proposed over the past decades of research and continue to be examined in the literature. Martin and Failows (2010) summarize the most frequently encountered theories of EF into five categories, namely the "narrowing" accounts, problem-solving account, "widening" accounts, "distributed" accounts, and the "unity in diversity" account.

"Narrowing" accounts. Theories that attempt to simplify the multidimensional view of EF into a single construct are referred to as "narrowing" accounts of EF (Martin & Failows, 2010). Firstly, this includes theories that emphasize a single aspect of EF as its central and defining feature. For example, Barkley (1997) emphasizes that inhibition is the central feature of EF. Additionally, Roberts and Pennington (1996) highlight that the central feature of EF is the *interaction* between inhibition and working memory. However, these views are generally considered inadequate to capture the substantial variability and complexity of EF abilities. Secondly, rather than focusing on a sole component of EF, some theories have proposed that EF can be entirely captured as a unitary construct. In this vein, some researchers have proposed that EF may operate as a "g" or general factor that requires many cognitive processes to operate successfully. Due to the substantial factor analytic work that has been done in EF, this narrowing view is also generally considered an oversimplification of EF (Lehto, Juujarvi, Kooistra & Pulkkinen, 2003; Messer et al., 2018).

Problem-solving account. According to the problem-solving account, EF is viewed as a function that can be defined in terms of its problem-solving outcomes (Martin & Failows, 2010). Zelazo and Müller's (2002) problem-solving framework specifies several stages of EF when solving a problem: problem representation, planning, execution (including intention and rule use) and evaluation (including error detection and correction). As such, the hierarchical structure and subfunctions of EF are all organized and conceptualized around the outcome of problem-solving.

“Widening” accounts. As opposed to “narrowing” accounts of EF, these accounts attempt to “widen” EF by focusing on its multiple components, which may be related to one another (Martin & Failows, 2010). According to this view, researchers aim to identify the various components of EF and subsequently investigate the relationships amongst them. For example, Hughes (2002) emphasizes the independence of components of EF, while acknowledging that the proposed relationships between EF abilities vary considerably.

“Distributed” accounts. Rather than viewing EF as a true “executive” function that oversees various components, this account posits that EF is located and spread across all of the associated components and processes (Martin & Failows, 2010). An important addition of the “distributed” accounts of EF is its potential amenability to include the external distribution of EF to the social and interpersonal world. This is noteworthy, as there is evidence to suggest that social interactions can significantly influence EF (Carpendale & Lewis, 2006). As such, the “distributed” view of EF acknowledges the distribution of EF both internally (i.e., cognitive or reasoning skills) and externally (i.e., social contexts).

“Unity in diversity” account. The “unity in diversity” view can be largely credited to Miyake et al. (2000), where they proposed their tripartite model of EF. Their latent variable analysis revealed that the three EFs selected (i.e., shifting, inhibition, and working memory or updating) were clearly distinguishable, though not completely independent and had some underlying shared commonality. Miyake et al. (2000) concluded that these three EFs demonstrated both unity and diversity, from where the title of these accounts is drawn. This work was later replicated by Lehto et al. (2003), for example, who obtained three similar factors that were considered separate, yet interrelated, in concordance with Miyake et al.’s (2000) model.

More recently, Cirino et al. (2018) investigated a framework of EF specific to children in the late elementary years. They delineated between three current approaches to conceptualizing EF: (1) functions related to EF that are disrupted by damage to the frontal lobes, (2) a collection of separate but related cognitive abilities, akin to Miyake et al's (2000) model, and (3) the use of EFs across stages of problem solving, derived from work in developmental psychology, neuropsychology, and educational psychology. By bridging across all of these literatures, they identified eight EF components: (1) working memory, (2) inhibition, (3) shifting, (4) planning, (5) generative fluency, (6) self-regulated learning, (7) metacognition, and (8) behavioural regulation. From their factor analytic work, a bifactor model with a common EF factor and five specific EF factors (i.e., working memory-span/manipulation and planning, working memory-updating, generative fluency, self-regulated learning and metacognition) best fit this data for these late elementary children. As such, this continues to provide evidence of EF as a multidimensional construct with overlapping and separable underlying components.

Development of Executive Functions

Like many other core cognitive abilities, EF abilities gradually change and improve over the course of development and can continue to ameliorate into adulthood (Best & Miller, 2011). EF abilities develop rapidly during the preschool years, and optimal performance can often be achieved in adolescence or early adulthood (Anderson, 2002; Steinberg, 2005). Throughout development, the prefrontal cortex (PFC) gains network efficiency, which supports the development of high-level cognitive processes, such as EFs (Bunge & Zelazo, 2006; Moriguchi & Hiraki, 2009). There has been some evidence to suggest that various EF components develop at different rates throughout childhood (Brod, Bunge & Shing, 2017, Huizinga, Dolan & van der Molen, 2006; Xu et al., 2013). For example, Davidson, Amso, Anderson and Diamond (2006) investigated the development of EFs in a sample of children aged 4 to 13. They found that even the youngest children in this sample were able to successfully exhibit EF skills related to working memory (i.e., hold information in their minds) and inhibition (i.e., inhibit a dominant response), as well as combine these skills in a single task. However, EF skills relating to cognitive flexibility (i.e., switching between rules) showed a much longer developmental progression, with even the 13-year old children not yet achieving adult-level performance.

It is also important to note that there is substantial evidence that EF abilities have important relationships to many key aspects of life, including mental and physical health, quality of life, school readiness, school success, job success, marital harmony, and public safety (Diamond, 2013). This remains true across development, as EFs have important correlates to crucial abilities in children (Arffa, 2007; Diamantopoulou, Rydell, Thorell & Bohlin, 2007). EF abilities are thought to be necessary for fostering successful development in children in a number of key domains including positive peer relations, intellectual abilities, and academic abilities (Arffa, 2007; Diamantopoulou et al., 2007). Given these important roles of EF, understanding the full spectrum of EF abilities is essential for all children, including those with and without developmental challenges.

EFs are often an area of focus within several neurodevelopmental conditions, including Attention-Deficit/Hyperactivity Disorder (ADHD). ADHD is defined as a neurodevelopmental disorder characterized by hyperactivity, inattention, and impulsivity (APA, 2013). Reliable differences have been found between individuals with ADHD and without ADHD on various EF measures, with EF deficits consistently found in children and adults with ADHD (Adler et al., 2017; Mahone et al., 2002; McLuckie et al., 2018; Willcutt et al., 2005). These EF deficits in children with ADHD are most prominently found in domains such as inhibition, working memory, planning, vigilance (Willcutt et al., 2005), and behaviour in familiar environments, such as at school or at home (Barkley & Fischer, 2011). Deficits in EF for children with ADHD have been associated with poor outcomes in adulthood (Barkley & Murphy, 2010). Results from neuroimaging studies have also supported findings regarding EF deficits in children with ADHD. For example, specific brain regions that are related to EF processes (e.g., frontostriatal and frontoparietal networks) have demonstrated underactivity, delay in cortical maturation, and decreased volume in individuals with ADHD (Faraone et al., 2015; Nigg, 2009; Shaw et al., 2007).

Assessing Executive Functions

There are currently two common methods used to assess EFs in children. First, EFs can be assessed using performance-based measures (Willcutt, Doyle, Nigg, Faraone & Pennington, 2005). Here, the child completes the task under the close supervision and direction of an examiner. Tasks are administered according to highly standardized procedures in a structured and optimal environment. For instance, this includes the stimuli being presented in a consistent manner, detailed instructions, prompts and feedback from the examiner, individual administration and a quiet and distraction-free testing room. The outcome variables in these measures are typically constructs such as accuracy or response time (Toplak et al., 2013). Some performance-based EF measures include the Stroop-Colour Word Test (Golden, 1978), the Trail-Making Test (Reitan, 1971), and the Wisconsin Card Sorting Task (Heaton, Chelune, Talley, Kay & Curtis, 1993).

Many of the classic performance-based tasks of EF aim to assess one of the core EF components identified in Miyake et al.'s (2000) tripartite theory of EF, namely shifting, inhibition, and working memory. Shifting abilities are often assessed through set-shifting tasks, which requires the individual to switch attention between tasks. The *Wisconsin Card Sorting Task* is a set-shifting task where the participant is told to sort cards based on different possible rules related to factors such as color, shape, or number. They are not told how to match the cards (i.e., the rule chosen by the examiner) and are only told when the match is wrong. The rule is then switched to a new deck. Perseveration errors are used as a dependent measure, since participants will often use a previous rule that is no longer valid despite the feedback (Heaton, Chelune, Talley, Kay & Curtis, 1993). The *Trailmaking Test* (part B) is another set-shifting task, in which the participants are presented with a series of circles containing both letters and numbers that they must connect, in order, as quickly as possible. They must switch between letters and numbers in order (e.g., A – 1- B – 2 – 3 – C, etc.) (Reitan, 1971).

A classic performance-based task of inhibition is the *Stop-Signal Task*. In these tasks, the participant performs a “go” task, such as pressing the left or right arrow keys based on the direction of the arrows presented on the screen. Occasionally the “go” stimulus (i.e., the arrow presented), is followed by a “stop signal” stimulus (i.e., an auditory beep), which instructs the individual to withhold their response. This “stop signal” stimulus is varied in time when it is paired with the “go” signals, which makes it more difficult for subjects to inhibit their responses (Verbruggen & Logan, 2008). Inhibition can also be assessed by measuring interference control*,* which is the ability to pay selective attention to a task with distracting or interfering information. The *Stroop Test* is a classic interference control task involving colors. While there are different variations of the test, often a participant is timed naming different colors on an assortment of color blocks provided. Then, in the interference condition, the participant receives a list of names of colors that are printed in a color incongruent with the word itself (e.g., the word “red” would be printed in green ink). In this condition, the participant must inhibit the name of the colour that is written, and only name the color of the ink (Golden, 1978).

Working memory is often evaluated in two ways: through auditory working memory and visual-spatial working memory (Diamond, 2013). Auditory working memory can be measured through the *Digit Span Task* which is a subtest on the Wechsler Intelligence Scale for Children

(WISC). In this task, a participant attempts to remember digits and repeat them in the reverse order in which they are presented or in numerical order, depending on the condition. Similarly, the *Letter-Number Sequencing Task* (also from the WISC) requires a participant to remember a string of letters and numbers together but must repeat the numbers in chronological order, and the letters in alphabetical order (Wechsler, 2014). Visual-spatial working memory can be measured through the *N-back Test*, for example. Here, a participant has to track letters that are presented consecutively on the computer screen and decide for each successive letter whether it was presented one (1-back), and in another condition, two (2-back) positions before (Wilhelm, Hildebrandt & Oberauer, 2013).

Second, EFs can be measured using behavioural ratings of EF abilities. These measures were developed to assess EFs in a more ecologically relevant manner than performance-based measures of EF (Roth, Isquith & Gioia, 2005). Here, a person who is very familiar with the child is presented with a list of behaviours related to cognitive and behavioural regulation, and the person is asked to rate the frequency and/or severity of the child's difficulties in these domains. For children, a secondary observer, such as a teacher or a parent, will typically complete these measures retroactively (Toplak et al., 2013). Some commonly used behavioural rating measures of EF include the Behavior Rating Inventory of Executive Function (BRIEF; Gioia, Isquith, Guy & Kenworthy, 2000), the Childhood Executive Functioning Inventory (CHEXI; Thorell, Eninger, Brocki & Bohlin, 2010), and the Barkley Deficits in Executive Functioning for Children and Adolescents (BDEFS-CA; Barkley, 2012).

Issues of Measurement of Executive Functions

As both types of measures are intended to assess EF abilities, one would expect that they would be highly correlated with one another from the perspective of convergent validity (Gregory, 2011). To the contrary, several studies over the past decade have reported only low to modest associations between performance-based and behavioural rating measures of EF in both typically-developing (TD) children and clinical samples, such as those with ADHD (Bodnar, Prahme, Cutting, Denckla & Mahone, 2007; Gray, Fettes, Woltering, Mawjee & Tannock, 2016; Mahone et al., 2002; McAuley, Chen, Goos, Schachar & Crosbie, 2010). This evidence suggests that these two EF measures are either not measuring the same construct or are not doing so in a consistent manner. This lack of convergent validity poses significant problems to clinicians who are assessing EF difficulties in children. For example, some children may be misidentified as having EF problems while others' EF difficulties may be entirely missed, depending on how EF is being measured.

Toplak, West and Stanovich (2013) examined 20 studies that investigated the issue of non-convergence of EF measures in a variety of samples. Similar to the studies mentioned above, Toplak et al. (2013) found a lack of concordance between performance-based and behavioural rating measures. In fact, only 24% of the correlations reported between measures of EF were significant, with a median correlation of .19. As such, these measures should not be viewed as interchangeable and they likely assess related but not identical constructs (Toplak et al., 2013).

Martin and Failows (2010) expand upon some of these methodological concerns regarding performance-based measures of EF. They explain that there is wide criticism regarding EF tasks due to their lack of construct validity, low test-retest reliability, lack of discriminant validity, and task impurity (i.e, any EF being assessed will likely implicate others at the same time, creating an impure task) (Brocki & Bohlin, 2004; Hughes, 2002; Miyake et al., 2000). As such, researchers often disagree on which EF component a given task is "purely" measuring. For example, Zelano and Müller (2002) have described the Tower of Hanoi task as a pure measure of planning, while Miyake et al. (2000) argue that it is more likely a measure of inhibition.

Task Structure and Self-Direction

It has been argued that the highly structured nature of the performance-based tasks of EF reduces the amount of self-direction required of the child, which is not reflective of the actual environments that children must navigate on a daily basis (Toplak et al., 2013). Thus, in these types of tasks, the structure imposed by the examiner greatly reduces the necessity for children to use their own self-directional skills (Salthouse, Atkinson & Berish, 2003). These skills may include deciding how to proceed when answering questions, completing tasks with little guidance or feedback, selecting strategies for proceeding, and deciding how to modify behaviour if needed. These types of skills are necessary in environments such as the home, school, or interacting with peers.

Some models of EF that stem from the clinical literature have emphasized the central role that self-direction plays in children's EF. For example, Barkley's model of EF highlights the importance of self-directed behaviour to help explain these difficulties in children with ADHD. This model elaborates that successful EF-related behavior requires the individual to employ various self-regulation processes (Barkley, 2012). Additionally, some conceptualizations of EF distinguish between external EF and self-directed EF (Snyder & Munakata, 2010). Under this distinction, EF can be externally driven (i.e., a child performs goal-directed behaviors with external reminders) or self-directed (i.e., a child must independently perform goal-directed behaviors without any external reminders).

Structure is documented as having a very important role for children, especially for those with ADHD (Imeraj et al., 2016; Mahone & Hoffman, 2007). A highly structured environment can provide support to children with ADHD and increase their attention span (Mahone & Hoffman, 2007), whereas unstructured environments cause greater difficulties for these children by reducing their attention span and ability to focus (Imeraj et al., 2016). Thus, traditional performance-based EF measures may not be ecologically relevant and potentially fail to deliver a clear picture of the EF issues experienced by children, perhaps due in part to the element of structure.

Unstructured Performance Task of Executive Function

To address the issue of structure in a performance-based EF task, Ledochowski, Andrade and Toplak (2019) developed the Unstructured Performance Task (UPT) of EF. This task was developed to minimize the amount of structure imposed by the examiner, thus increasing the self-direction required of the child with the aim of resembling more ecologically valid tasks that require the use of executive function in everyday activities for children. Children were presented with 42 easy language or math problems scattered randomly on a large sheet of paper. They were not given an explicit time limit, detailed instructions, or prompts from the examiner to stay on task or continue the task. In this way, Ledochowski et al. (2019) hoped to mimic daily activities of children that are presented with this lack of structure. In fact, the UPT was developed with the intention of resembling many typical school activities that children would have to complete on a daily basis, such as work sheets. The unstructured nature of the UPT was intended to require children to engage in their own self-directional behaviours, such as monitoring completion and organizing strategies to accurately complete questions.

Ledochowski et al. (2019) piloted the UPT in samples of 8 to 12-year-old children with ADHD (*n*=38) and a community sample of children (*n*=42). As predicted, TD children performed significantly better on the UPT than children with ADHD. They also rated their subjective performance of the task differently, with children with ADHD rating the task as more difficult. Additionally, they reported investing a significantly lower degree of cognitive effort than the TD group. Traditional performance-based tasks of EF and behavioural ratings of EF were also administered, and they were not significantly correlated with each other. However, the

UPT was significantly correlated with each type of EF measure. The UPT also predicted ADHD status more accurately than traditional performance-based measures of EF. In contrast, the UPT predicted ADHD status similarly to behavioural ratings. However, as item difficulty was designed to be relatively low, TD children generally displayed a ceiling performance in the UPT, with a large proportion of children achieving a perfect or a near-perfect score. Nonetheless, the UPT generated promising findings in this pilot study. The next step in this line of research that is needed is to further investigate and refine the UPT to increase its utility in both TD and ADHD populations.

Present Study

In the present project, two studies were conducted. In Study 1, the psychometric characteristics of the original UPT were further investigated based on collected data from two samples of children on this task. Specifically, reliability, test structure, and key correlates of the UPT were examined in depth to better understand the psychometric properties of the UPT. From these findings, various elements of the UPT were further developed with the aim of generating greater variability on this task in a community sample of children, and of improving its psychometric properties. In Study 2, the updated UPT was piloted in a community sample of children. In this study, the UPT-2 was administered alongside traditional performance-based and behavioural ratings of EF. In addition, measures of academic abilities were administered to assess associations with performance on the UPT-2. This project comprised a critical step in the development of a performance-based measure of EF that requires self-direction from examinees to complete this task.

Research Objectives

This project had three primary objectives. The first aim was to examine the psychometric characteristics (i.e. internal consistency and split-half reliability) and underlying task structure (i.e. math, language, and symbolic domains) of the UPT (Study 1a). The second aim was to examine the pattern of associations between behavioural ratings of EF, performance-based measures of EF, academic abilities (i.e. reading and math), ratings of emotional and behavioural, age and the UPT (Study 1b). The third aim was to develop and pilot an updated version of the UPT (i.e. UPT-2) and to examine its psychometric properties and correlates to other key measures (Study 2).

Hypotheses

In order to address the three primary research objectives as stated above, there were three hypotheses.

Hypothesis 1. Overall, it was hypothesized that the UPT would demonstrate good psychometric properties. Specifically, the UPT would demonstrate good internal consistency and split-half reliability. When examining task structure (i.e. math and language domains), performance on these domains and correlates would not differ significantly, suggesting that the UPT is measuring an overall construct rather than math and language abilities separately.

Hypothesis 2. The UPT was hypothesized to be significantly correlated to behavioural ratings of EF, performance-based measures of EF, academic abilities (i.e. reading and math), difficulties in emotional and behavioural functioning, and age. However, the strength of these relationships would not significantly differ across the math and language domains of the UPT.

Hypothesis 3. The UPT-2 was hypothesized to generate more variability in scores in a community sample than the original UPT, minimizing the presence of a ceiling effect. The UPT-2 would also have good psychometric properties and display expected relationships to other measures.

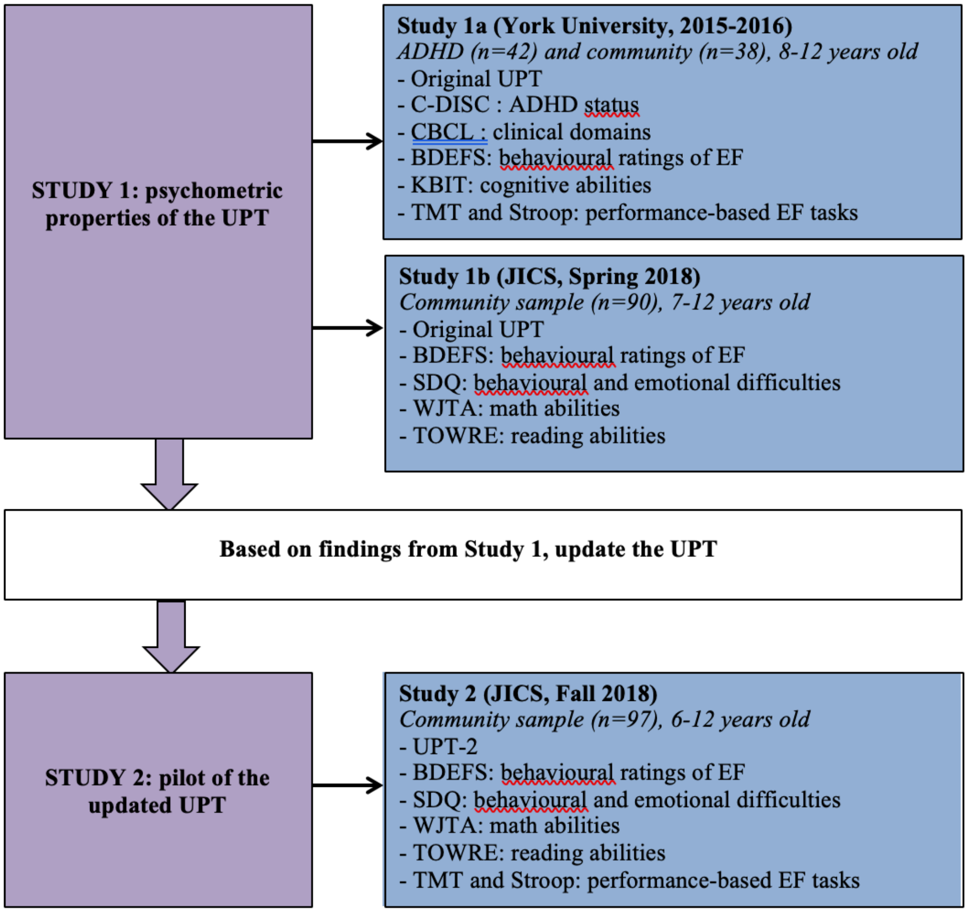

STUDY 1: psychometric properties of the UPT
Study 1a (York University, 2015-2016)
ADHD (n=42) and community (n=38), 8-12 years old
- Original UPT
- C-DISC : ADHD status
- CBCL : clinical domains
- BDEFS: behavioural ratings of EF
- KBIT: cognitive abilities
- TMT and Stroop: performance-based EF tasks
Study 1b (JICS, Spring 2018)
Community sample (n=90), 7-12 years old
- Original UPT
- BDEFS: behavioural ratings of EF
- SDQ: behavioural and emotional difficulties
- WJTA: math abilities
- TOWRE: reading abilities
Based on findings from Study 1, update the UPT
STUDY 2: pilot of the updated UPT
Study 2 (JICS, Fall 2018)
Community sample (n=97), 6-12 years old
- UPT-2
- BDEFS: behavioural ratings of EF
- SDQ: behavioural and emotional difficulties
- WJTA: math abilities
- TOWRE: reading abilities
- TMT and Stroop: performance-based EF tasks

Figure 1. Summary of the methodology and measures across studies.

Methods

Participants

Study 1a. Eight-six children were recruited from the Greater Toronto Area to participate in Study 1a, which took place at York University. This sample of participants consisted of two groups of children: a community sample of children and a clinical sample of children with diagnosed ADHD. The community sample consisted of 15 females and 27 males aged 8-12 (*M*=9.57, *SD*=1.23). The ADHD group consisted of 11 females and 27 males aged 8-12 (*M*=9.55, *SD*=1.37). These groups did not significantly differ in terms of gender (*p*=.52) or age (*p*=.82). Please refer to Ledochowski et al. (2019) for further details on this sample, including recruitment, inclusion/exclusion criteria, medication status, ethnicity, and comorbidities.

Study 1b. In partnership with collaborators at the Ontario Institute for Studies in Education (OISE, University of Toronto), a community sample of children in grades 2-6 was recruited from the Dr. Eric Jackman Institute of Child Study lab school (JICS) in May 2018, as a part of an ongoing longitudinal study. Ninety-one children from JICS participated in this study. However, one child in grade 2 was excluded from this study as they were unable to complete the UPT without a scribe. Thus, a total of 90 children between the ages of 7-12 (*M*=10.01, *SD*=1.45) were included in the analyses. This sample consisted of 48 boys (M_{age}=10.02, *SD*=1.4) and 42 girls (M_{age}=10.01, SD=1.52). There were 17 participants in grade 2 (M_{age}=7.96, *SD*=0.27, 8 girls), 17 in grade 3 (M_{age}=8.95, *SD*=0.27, 10 girls), 17 in grade 4 (M_{age}=9.87, *SD*=0.27, 9 girls), 20 in grade 5 (M_{age}=10.96, *SD*=0.32, 11 girls), and 19 in grade 6 (M_{age}=11.94, *SD*=0.28). Five children were reported as having diagnosed ADHD, 10 children were reported as having a diagnosed LD, and one child was reported as having diagnosed ASD. There was missing parent questionnaire data for seventeen children in this sample.

Study 2. In partnership with collaborators at the Ontario Institute for Studies in Education (OISE, University of Toronto), a community sample of children in grades 1-6 was recruited from the Dr. Eric Jackman Institute of Child Study lab school (JICS) in October 2018, as a part of an ongoing longitudinal study. Ninety-eight children from JICS participated in this study. However, one child in grade 1 was excluded as she/he was unable to complete the UPT. Thus, a total of 97 children between the ages of 5.23 and 11.92 (*M*=9.09, *SD*=1.76) were included in the analyses. This sample consisted of 51 boys (M_{age}=9.17, *SD*=1.74) and 46 girls (M_{age}=9.01, *SD*=1.80). There were 12 participants in grade 1 (M_{age}=6.14, *SD*=0.36), 17 participants in grade 2 (M_{age}=7.55, *SD*=0.29), 18 participants in grade 3 (M_{age}=8.52, *SD*=0.27), 16 participants in grade 4 (M_{age}=9.47, *SD*=0.27), 14 participants in grade 5 (M_{age}=10.39, *SD*=0.29), and 20 participants in grade 6 (M_{age}=11.47, *SD*=0.32). Four children were reported as having diagnosed ADHD and seven children were reported as having a diagnosed LD. There was missing parent questionnaire data for fourteen children in this sample.

Procedure

Study 1a. This study was embedded within a larger study at York University from 2015-2016 that examined performance calibration. Two examiners met with each child and their parent(s). The study was explained to each child and parent separately, and informed consent and assent was obtained. One examiner administered measures to the child while another examiner administered measures to the parent. Time of testing ranged from 90 to 120 minutes, and each participant received a small honorarium of $20 upon completion.

Study 1b. This study was embedded within a larger longitudinal study at the Dr. Eric Jackman Institute for Child Studies (JICS), consisting of the first time point of data collection in Spring 2018. For this study, all children in grades 2 to 6 were invited to participate in this study by receiving a parental consent form to bring home to their parents. Once parental consent was received, the child was met at school by an examiner to complete informed assent. They were then accompanied to a quiet testing room at JICS to complete the necessary components of this study. Time of testing ranged from 60 to 90 minutes.

Study 2. This study was also embedded within the same larger longitudinal study at JICS, consisting of the second time point of data collection in Fall 2018. For this study, children in grades 1 to 6 were invited to participate. For returning participants, parents were sent a reminder email about data collection with the option to withdraw and all children completed a re-assent process with an examiner. New participants were also invited to participate (i.e. younger students who were now old enough to participate or new students at JICS) by receiving a parental consent form. If their parent agreed to participate, then the child completed the assent process with an examiner. They were then accompanied to a quiet testing room at JICS to complete the necessary components of this study. Time of testing ranged from 60 to 90 minutes.

Measures

Experimental measure

Unstructured Performance Task of Executive Functions (UPT; Ledochowski et al., 2019). The UPT is a novel performance-based measure of EF for children. The task is presented on a large sheet of paper (11x17 inches), and contains simple questions in the domains of math, reading, general knowledge, and rote copying. Questions are presented in a random, unstructured order on the page and are not numbered, leaving it up to the participant to decide how to approach this task. Additionally, the children are given very few instructions or prompts from the examiner and cannot be helped with the task. In Study 1a and 1b the original UPT (42 items, one sided, approximately 10 minutes; Appendix A) was administered and in Study 2 the UPT-2 (50 items, double sided, approximately 15 minutes; Appendix B) was administered. The same brief instructions were presented to each participant: "I would like you to complete the following worksheet. If you do not know the answer for any of the problems, just circle it and go on to the next problem. I cannot read any of the questions to you. Just do your very best, and when you are done, please bring the worksheet to me." In both versions, the UPT was scored on correctness (i.e. was the response to the item correct) and completion (i.e. was the item attempted or circled, regardless of accuracy).

Study 1a.

Demographics Questionnaire. Parents completed a brief demographics questionnaire that provided information about themselves and their child. Questions included information such as the child's age, gender, country of birth, ethnicity and medication as well as parental ethnicity, marital status, educational status, and occupational status.

Computerized Diagnostic Interview Schedule for Children – Parent Version (C-DISC; Fisher et al., 2006). The C-DISC is a computerized structured interview designed to assess DSM-IV psychiatric disorders, symptoms, and level of impairment in children and adolescents aged 6 to 17. Trained graduate students administered the Attention/Deficit-Hyperactivity Disorder, the Oppositional Defiant Disorder, and the Conduct Disorder subscales. This measure was used to characterize each sample and to confirm ADHD diagnosis.

Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001). The CBCL is a checklist completed by parents to assess emotional and behavioural problems in youth aged 6 to 18. This measure generates six DSM-5 oriented scales which include affective problems, anxiety problems, somatic problems, ADHD, oppositional defiant problems, and conduct problems.

Barkley Deficits in Executive Functioning Scale – Children and Adolescents Short Form (BDEFS-CA; Barkley, 2012). The BDEFS-CA Short Form is a parent questionnaire used to assess EF in children and adolescent's daily activities. As such, it contains questions pertaining to domains of time management, problem solving, organization, self-restraint, self-motivation, and self-regulation of emotions. The short form contains 20 items, which are rated as 1=Never, 2=Sometimes, 3=Often, and 4=Always, where parents rate their child's behaviour over the past six months. The EF Summary Score was used, which is the total score comprised by summing the answer of all 20 questions, with a possible range of scores from 20-80.

Kaufman Brief Intelligence Test, Second Edition (KBIT-2; Kaufman & Kaufman, 2004). The KBIT-2 is a brief measure of crystallized and fluid intelligence. Crystalized intelligence was assessed with two orally presented verbal tasks (receptive and expressive vocabulary tasks) that do not involve reading or spelling, and ask questions pertaining to verbal knowledge and riddles. Fluid intelligence was assessed non-verbally with a matrix-reasoning task. The raw scores of each of these subtests were standardized and summed to create a non-age corrected *z*-score.

Trail-Making Test (TMT; Reitan, 1971). The TMT is a performance-based measure of EF, that specifically assesses set-shifting. Set-shifting is defined as a cognitive task that requires one to display flexibility when there are changing rules or schedules of reinforcement in their environment (Strauss et al., 2006). This task is administered by pencil and paper and guided by the examiner. In Part A of this task, participants are asked to connect 25 numbered circles in numerical order using a pencil. In Part B of this task, participants were asked to connect alternating letters and numbers in alpha-numerical order (i.e. 1 to A, A to 2, 2 to B, B to 3, etc.). In this part there were 13 numbers and 12 letters. The outcome measure for this task was calculated by subtracting completion time on Part A from completion time on Part B. As such, this controls for processing speed, and generates a score of set-shifting.

Stroop Color-Word Test (Golden, 1978). The Stroop is another performance-based measure of EF, that specifically assesses interference control. Interference control is a type of inhibition that is defined as the ability to filter out irrelevant information and select relevant information. There were two conditions, each containing 48 items arranged in a 6x8 matrix. In the colour naming condition, participants were presented with 48 patches of colour (red, blue, green or yellow), and asked to name the colours as quickly as possible without making any errors. In the interference condition, participants were presented with 48 words (RED, BLUE, GREEN or YELLOW) that were printed in an incongruent ink colour (e.g. the word red is printed in yellow). Participants were asked to name the colour in which the word was printed as quickly as possible without making any errors. The outcome measure for this task was calculated by subtracting the total time on the colour naming condition from the total time on the interference condition, which provides the inhibition score (Strauss et al., 2006).

Study 1b.

Demographics Questionnaire. Parents completed a brief demographics questionnaire that provided information about themselves and their child. Questions included information such as the child's age, gender, languages spoken, neurodevelopmental diagnoses and medication as well as parental age and educational status.

Barkley Deficits in Executive Functioning Scale – Children and Adolescents Short Form (BDEFS-CA-CA; Barkley, 2012). The BDEFS-CA-CA Short Form is a parent questionnaire used to assess EF in children and adolescent's daily activities (please see above for further details).

Strengths and Difficulties Questionnaire – Parent Version (SDQ; Goodman, 1997). The SDQ is a parent questionnaire used to assess emotional and behavioural difficulties in children and adolescents. As such, it contains questions pertaining to conduct problems, emotional problems, ADHD symptoms, peer problems, and prosocial behaviour. Parents were provided with the appropriate version of the questionnaire according to their child's age (i.e. ages 4-10 or ages 11-17). This scale contains 25 items, which are rated as 1=Not True, 2=Sometimes True, 3=Certainly True. A total difficulties score was generated by summing all items (range of 25-75), as well as domain scores of conduct problems (range of 5-15), emotional difficulties (range of 5-15), ADHD symptoms (range of 5-15), peer difficulties (range of 5-15), and prosocial behaviour (range of 5-15).

Woodcock-Johnshon Tests of Achievement – Fourth Edition (WJTA-IV; Schrank et al., 2014). The WJTA-IV is a standardized assessment measure of academic achievement commonly used in children and adolescents. Participants were administered two math subtests: Math Fluency and Math Calculation. Math Fluency is a timed task where participants were asked to solve as many simple math problems (i.e. addition, subtraction, multiplication) as possible in three minutes. Math Calculation is not a timed task where participants were asked to solve as many math problems as they could of increasing difficulty.

Test of Word Reading Efficiency – Second Edition (TOWRE-2; Torgesen et al., 2011). The TOWRE-2 is a standardized measure that assesses efficiency of sight word recognition and phonemic decoding in children and adults. The Sight Word Reading Efficiency and Phonemic Decoding subtests were administered to participants in this study. Sight Word Reading Efficiency assesses the number of real printed words that participants accurately read within 45 seconds. Phonemic Decoding measures the number of pronounceable and printed non-words that participants accurately decoded within 45 seconds.

Study 2.

Demographics Questionnaire. Parents completed a brief demographics questionnaire that provided information about themselves and their child (please see above for more details).

Barkley Deficits in Executive Functioning Scale – Children and Adolescents Short Form (BDEFS-CA-CA; Barkley, 2012). The BDEFS-CA-CA Short Form is a parent questionnaire used to assess EF in children and adolescent's daily activities (please see above for further details).

Strengths and Difficulties Questionnaire – Parent Version (SDQ; Goodman, 1997). The SDQ is a parent questionnaire used to assess emotional and behavioural difficulties in children and adolescents (please see above for further details).

Woodcock-Johnshon Tests of Achievement – Fourth Edition (WJTA-IV; Schrank et al., 2014). The WJTA-IV is a standardized assessment measure of academic achievement commonly used in children and adolescents. Participants were administered two math subtests: Math Fluency and Math Calculation (please see above for more details).

Test of Word Reading Efficiency – Second Edition (TOWRE-2; Torgesen et al., 2011). The TOWRE-2 is a standardized measure that assesses efficiency of sight word recognition and phonemic decoding in children and adults. The Sight Word Reading Efficiency and Phonemic Decoding subtests were administered to participants in this study (please see above for more details).

Trail-Making Test (TMT; Reitan, 1971). The TMT is a performance-based measure of EF, that specifically assesses set-shifting (please see above for more details).

Stroop Color-Word Test (Golden, 1978). The Stroop is another performance-based measure of EF, that specifically assesses interference control (please see above for more details).

Statistical analyses

All statistical analyses were conducted using R. The significance level throughout this project was set at a standard of $p < .05$. Normality of each of the variables was tested using the Shapiro-Wilks test of normality as well as visual inspection of histograms and Q-Q plots. Some of the distributions were not normal and therefore nonparametric tests such as Spearman Rank Order Correlations and Independent Samples Mann-Whitney U Tests were employed throughout. All assumptions for multiple and logistic regression were met.

Across studies, multiple comparisons were made, particularly with many correlations being run. Although typically these multiple comparisons would call for corrections to the alpha level to avoid Type I errors, we chose not to follow this procedure. Due to the novelty of the UPT, we took an exploratory approach to the analyses. Most methods for multiple testing (e.g., Bonferroni corrections) were designed for confirmatory data analysis, and have been found to be inappropriate for exploratory analyses (Goeman & Solari, 2011). This being said, we interpreted the current results with caution until further confirmatory work is undertaken on the UPT.

To examine performance on the UPT, the descriptive statistics and distributions of correct and complete items on the UPT were examined. Indices of internal consistency (Cronbach's alpha) and split-half reliability (Spearman-Brown's formula) were used to examine the psychometric properties of the UPT. Spearman correlations were used to assess the relationships between the UPT and all of the outcome measures due to concerns regarding normality. Hierarchical regression models were also used in order to understand the contributions of various outcome measures in predicting scores on the UPT. Additionally, binomial logistic regression models were also used to determine whether the UPT and other measures predicted clinical risk as measured by the SDQ.

Results

Study 1a

An examination of the variables used in this study was conducted in order to ensure that assumptions were met before proceeding with further analyses. Firstly, histograms and Q-Q plots for all variables were visually inspected. Indices of skewness and kurtosis were then calculated for all variables. Shapiro-Wilks tests of normality were also conducted to assess the normality of the distribution of all variables. Results from these analyses are summarized in Table 1.

Table 1

Means, standard deviations, skewness and kurtosis indices, and W-statistics for all variables

	Mean (SD)	**Skewness**	**Kurtosis**	**W-statistics**
UPT total correct	35.16 (7.19)	-1.89	3.05	0.75**
UPT total complete	39.24 (5.01)	-3.36	12.44	0.55**
KBIT verbal	60.96 (10.78)	-0.12	-0.12	0.99
KBIT non-verbal	31.57 (5.00)	-0.46	0.08	0.96*
CBCL depression	2.80 (3.79)	1.87	3.42	0.74**
CBCL anxiety	3.86 (4.05)	1.07	0.19	0.85**
CBCL somatic	1.08 (1.71)	2.83	12.19	0.66**
CBCL ADHD	5.24 (4.25)	0.25	-1.29	0.92**
CBCL ODP	3.10 (2.85)	0.63	-0.72	0.90**
CBCL CP	2.61 (3.17)	0.19	0.49	0.81**
Stroop (Interference)	52.21 (23.98)	0.84	0.58	0.93**
TMT (Part B minus Part A)	115.58 (68.88)	1.14	0.64	0.88**
BDEFS-CA	52.33 (10.87)	0.45	-1.15	0.91**

Note. * p<.05; ** p<.01. UPT=Unstructured Performance Task, KBIT=Kaufman Brief Intelligence Test, CBCL=Child Behavioral Checklist, TMT=Trail Making Test, BDEFS-CA=Barkley Deficits in Executive Functioning Scale.

These analyses revealed that many of the variables in this study did not meet the assumption of normality. When looking at the primary outcome variable in this study (i.e., the UPT), it becomes apparent that there are concerns regarding normality. This was expected for the UPT based on Ledochowski et al.'s (2019) findings regarding a ceiling effect on this task. As such, non-parametric analyses were used in this study, namely Mann-Whitney U tests as well as Spearman correlations.

Psychometric properties of the UPT. Reliability indices were calculated for the UPT in order to better understand the psychometric properties of this task. When examining the UPT across both groups, this task demonstrated good internal consistency (α=.92) and split-half reliability (*ρ*=.90). Internal consistency of the UPT was higher amongst children with ADHD (α=.93) than those in the control group (α=.73). When looking at the distribution of scores on the UPT for these groups (Figure 2 and Figure 3), it becomes apparent that the ADHD group has a greater spread of scores for correct items as well as complete items. In comparison, the control group has a very restricted range of scores, pointing towards a ceiling effect in this group. When looking at the UPT item by item, most items had a very high mean (*M*=0.80-0.95), which further supports a potential ceiling effect on this task. A few items with lower means (i.e. <.75) were identified such as: "Give a word that ends with the letter G" (*M*=.68), "Do pickle and bickle rhyme?" (*M*=.70), "Do bark and part rhyme?" (*M*=.59), "9x3" (*M*=.59), "3x4" (*M*=.66), "What rhymes with face?" (*M*=.72), and "How many letter t's are in this sentence: The turtle ate tulips." (*M*=.37).

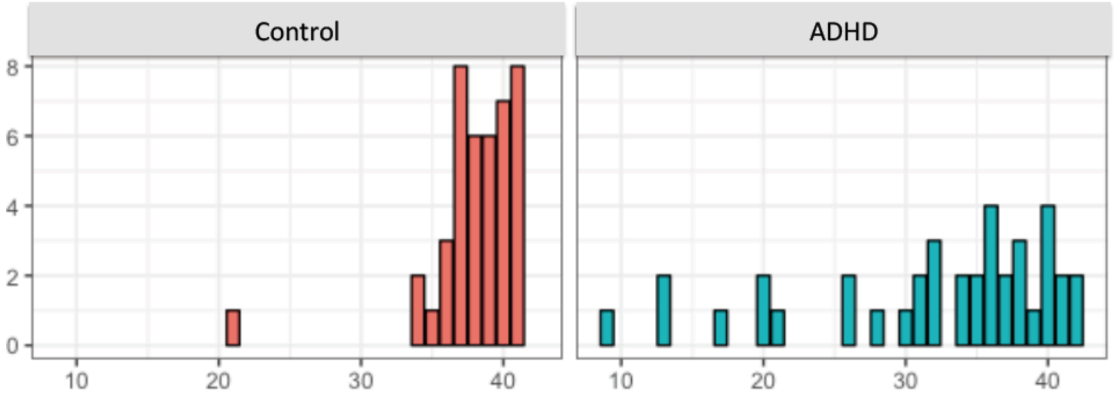


Figure 2. Distribution of correct items on the UPT by group.

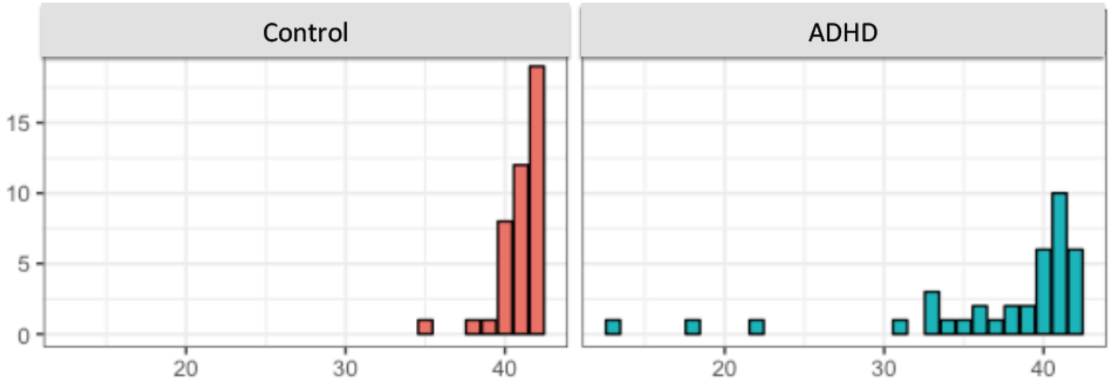


Figure 3. Distribution of complete items on the UPT by group.

Domains of the UPT. In order to better understand the task structure of the UPT, we divided the 42 items on the UPT into two domains, based on whether they were linguistically focused (i.e. language domain) or mathematically focused (i.e. math domain). After item-by-item examination of the UPT, 30 items were identified as belonging in the language domain and 12 items were identified as belonging in the math domain. Participants' performance did not significantly differ in the language (*M*=.84) and math domains (*M*=.82) in terms of correct items [*W*=3215, *p*=.85]. As expected, the control group (*M*=.91) had significantly more correct items than the ADHD group (*M*=.78) in the language domain [*W*=1114.5, *p*<.001]. The control group (*M*=.90) also had significantly more correct items than the ADHD group (*M*=.73) in the math domain [*W*=1108, *p*=.002] (Figure 4). When looking at completion of items, participants' performance also did not significantly differ in the language (*M*=.94) and math (*M*=.91) domains [*W*=3400.5, *p*=.37]. The control group (*M*=.98) completed significantly more items than the ADHD group (*M*=.90) in the language domain [*W*=1114, *p*<.001]. Additionally, the control group (*M*=.97) completed significantly more items than the ADHD group (M=.86) in the math domain [*W*=1044.5, *p*=.01].

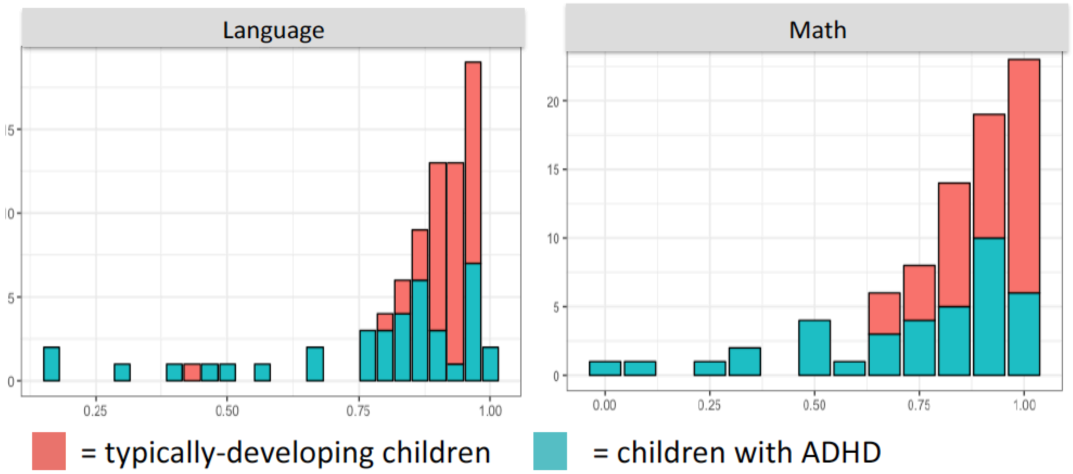


Figure 4. Distribution of correct items on the UPT language and math domains according to group.

Reliability indices were also calculated for these domains of the UPT. The language domain demonstrated good internal consistency (α=.91) when looking across the entire sample. The ADHD group (α=.93) demonstrated higher internal consistency than the control group (α=.72) in the language domain. This domain also demonstrated good split-half reliability (*ρ*=.91). The math domain also demonstrated good internal consistency (α=.81) when looking across the entire sample. However, when examining internal consistency within groups, the ADHD group (α=.85) continued to demonstrate good internal consistency in math, while the control group (α=.40) demonstrated poor internal consistency in math. Additionally, the math domain demonstrated moderate split-half reliability (*ρ*=.72).

Within the language domain, some of the items were identified as having strong pictorial content. As such, the language domain was further subdivided into pictorial language items (*n*=7) and non-pictorial language items (*n*=23). When looking across groups, participants performed better on pictorial language items (*M*=.91) than non-pictorial language items (*M*=.82) [*W*=4610, *p*<.001].

Correlates of the UPT. Relationships between the total UPT, language domain of the UPT and math domain of the UPT and several measures were calculated using Spearman correlations. Specifically, the UPT was correlated with verbal and non-verbal cognitive abilities (KBIT), clinical domains of depression, anxiety, somatic problems, ADHD, oppositional defiant problems and conduct disorder (CBCL), two performance-based measures of EF (Stroop and TMT) and one behavioural rating of EF (BDEFS-CA). Results from these correlations are summarized in Table 2.

Table 2

Spearman correlations between the UPT performance and cognitive abilities, clinical domains, and executive functions

		UPT total correct	**UPT language correct**	**UPT math correct**
KBIT	Verbal	.48**	.44**	.41**
	Non-verbal	.51**	.40**	.50**
CBCL (parent)	Depression	-.33**	-.31**	-.24*
	Anxiety	-.35**	-.30**	-.33**
	Somatic	-.15	-.04	-.23*
	ADHD	-.33**	-.30**	-.28*
	ODP	-.30**	-.30**	-.22*
	CP	-.27*	-.25*	-.20
EF	Stroop (Interference)	-.38**	-.36**	-.31**
	TMT (Part B minus Part A)	-.55**	-.49**	-.47**
	BDEFS-CA (parent)	-.36**	-.37**	-.27*

Note. * p<.05; ** p<.01. UPT=Unstructured Performance Task, CBCL=Child Behavioral Checklist, TMT=Trail Making Test, BDEFS-CA=Barkley Deficits in Executive Functioning Scale.

Generally, these relationships were significant across domains of the UPT. Tests of dependent correlations using Steiger's Z test were also calculated to determine whether any of the outcome variables were significantly more related to one domain of the UPT or the other. No correlation comparisons were found to be significant, suggesting fairly consistent associations across math and language domains.

Study 1b

An examination of the variables used in this study was conducted in order to ensure that assumptions were met before proceeding with further analyses. Firstly, histograms and Q-Q plots for all variables were visually inspected. Indices of skewness and kurtosis were then calculated for all variables. Shapiro-Wilks tests of normality were also conducted to assess the normality of the distribution of all variables. Results from these analyses are summarized in Table 3.

Table 3

Means, standard deviations, skewness and kurtosis indices, and W-statistics for all variables

	Mean (SD)	**Skewness**	**Kurtosis**	**W-statistics**
UPT total correct	37.92 (5.12)	-3.15	13.15	0.66**
UPT total complete	40.21 (4.19)	-4.58	24.31	0.44**
WJTA math fluency	63.69 (30.24)	1.17	1.40	0.90**
WJTA math calculation	27.78 (7.23)	0.21	-0.79	0.97
TOWRE phonemic decoding	40.74 (11.44)	-0.29	-0.75	0.97
TOWRE single word reading	71.16 (13.44)	-0.17	-0.05	0.99
SDQ total difficulties	6.66 (4.60)	0.75	0.60	0.95**
SDQ emotional problems	1.75 (2.01)	1.29	1.41	0.83**
SDQ conduct problems	0.96 (1.26)	1.58	2.53	0.76**
SDQ hyperactivity	3.03 (2.70)	0.55	-0.97	0.89**
SDQ peer problems	0.96 (1.42)	1.91	4.04	0.71**
SDQ prosocial behaviors	8.71 (1.57)	-1.10	0.15	0.80**
BDEFS-CA	34.28 (8.81)	0.49	-0.52	0.95**

Note. * p<.05; ** p<.01. UPT=Unstructured Performance Task, WJTA=Woodcock-Johnson Tests of Achievement, TOWRE=Test of Word Reading Efficiency, SDQ=Strengths and Difficulties Questionnaire, BDEFS-CA=Barkley Deficits in Executive Functioning Scale.

These analyses revealed that most of the variables in this study, with the exception of the measures of academic abilities (i.e., WJTA and TOWRE), did not meet the assumption of normality. When looking at the primary outcome variable in this study (i.e., the UPT), it becomes apparent that there are important concerns regarding normality. This was expected for the UPT based on Ledochowski et al.'s (2019) findings regarding a ceiling effect on this task, which was once again identified in Study 1a. As such, non-parametric analyses were used in this study, namely Kruskal-Wallis tests as well as Spearman correlations.

Performance on the UPT. Across the entire sample, participants answered an average of 39.77 (*SD*=5.93) items out of 42 correctly on the UPT. The number of correct items increased developmentally, according to grade level. As such, participants in grade two answered 33.12 (*SD*=8.15) items correctly, participants in grade three answered 36.53 (*SD*=4.96) items correctly, participants in grade four answered 38.65 (*SD*=3.02) items correctly, participants in grade five answered 39.95 (*SD*=1.47) items correctly, and participants in grade six answered 40.68 (*SD*=1.34) items correctly (Figure 5). A Kruskal-Wallis test revealed that there was a significant effect of grade on total correct items on the UPT [$\chi^2(4)=31.15$, $p<.001$]. Post-hoc analyses revealed that grades two and four, two and five, two and six, three and five, and three and six differed significantly, with participants in higher grades performing better than those lower grades. There were no significant differences between males (*M*=37.58, *SD*=5.77) and females (*M*=38.31, *SD*=4.31) on UPT total correct scores [*W*=1107.5, *p*=.42]. Cronbach's alpha for the total items correct on the UPT revealed good internal consistency (α=.88).

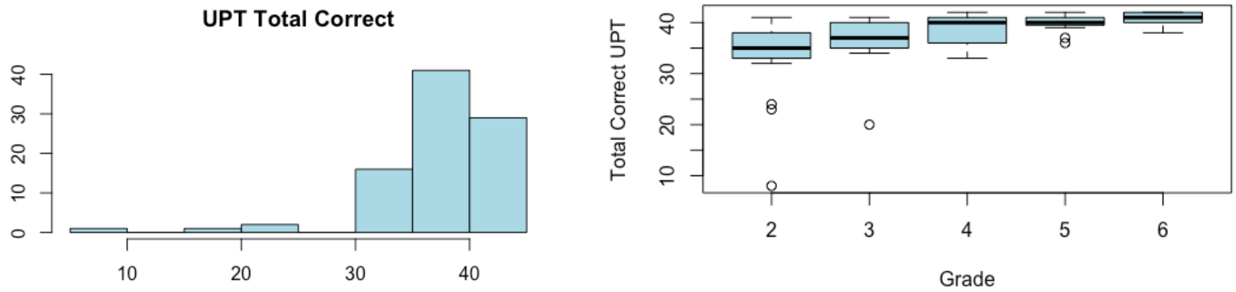


Figure 5. Distribution of correct items on the UPT across the entire sample and by grade.

Across the entire sample, participants completed an average of 40.21 (*SD*=4.19) items out of 42 on the UPT. The number of items completed increased developmentally, according to grade level. As such, participants in grade two completed 37.25 (*SD*=7.78) items, participants in grade three completed 39.49 (*SD*=4.50) items, participants in grade four completed 41.12 (*SD*=1.04) items, participants in grade five completed 41.15 (*SD*=1.04) items, and participants in grade six completed 41.53 (*SD*=0.90) items (Figure 6). A Kruskal-Wallis test revealed that there was a significant effect of grade on total complete items on the UPT [$\chi^2(4)=9.29$, *p*=.05]. Post-hoc analyses revealed that grades two and six differed significantly, indicating more completed items in higher grades. There was no significant difference between males (*M*=39.94, *SD*=5.02) and females (*M*=40.52, *SD*=3.00) on UPT total complete scores [*W*=1170.5, *p*=.16]. Cronbach's alpha for the total items complete on the UPT revealed good internal consistency (α=.93).

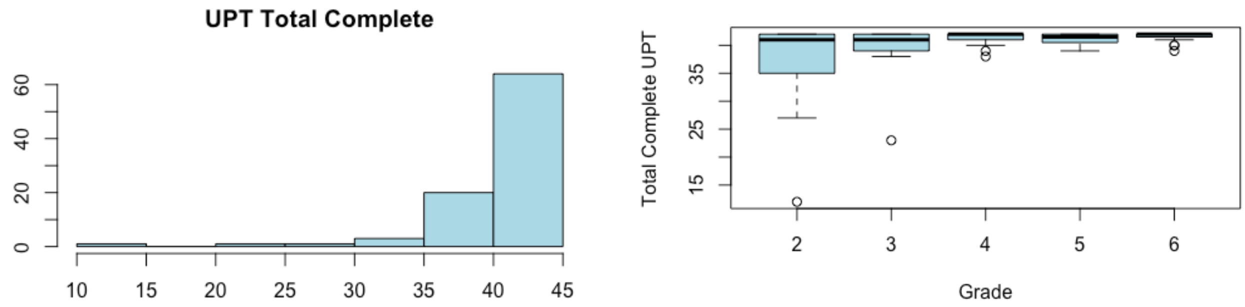


Figure 6. Distribution of complete items on the UPT across the entire sample and by grade.

As identified in Study 1a, the UPT was examined in terms of items from the math domain (*n*=12) and the language domain (*n*=30). Across the entire sample, participants completed 88% of math items correctly and 92% of language items correctly. Participants' performance in these two domains did not differ significantly [*W*=4527, *p*=.09]. Additionally, the developmental trends of the total UPT scores identified, such that performance increases with grade level, were consistent across the math and language domains of the UPT (Figure 7). Cronbach's alpha for the math domain (α=.84) and the language domain (α=.81) revealed good internal consistency.

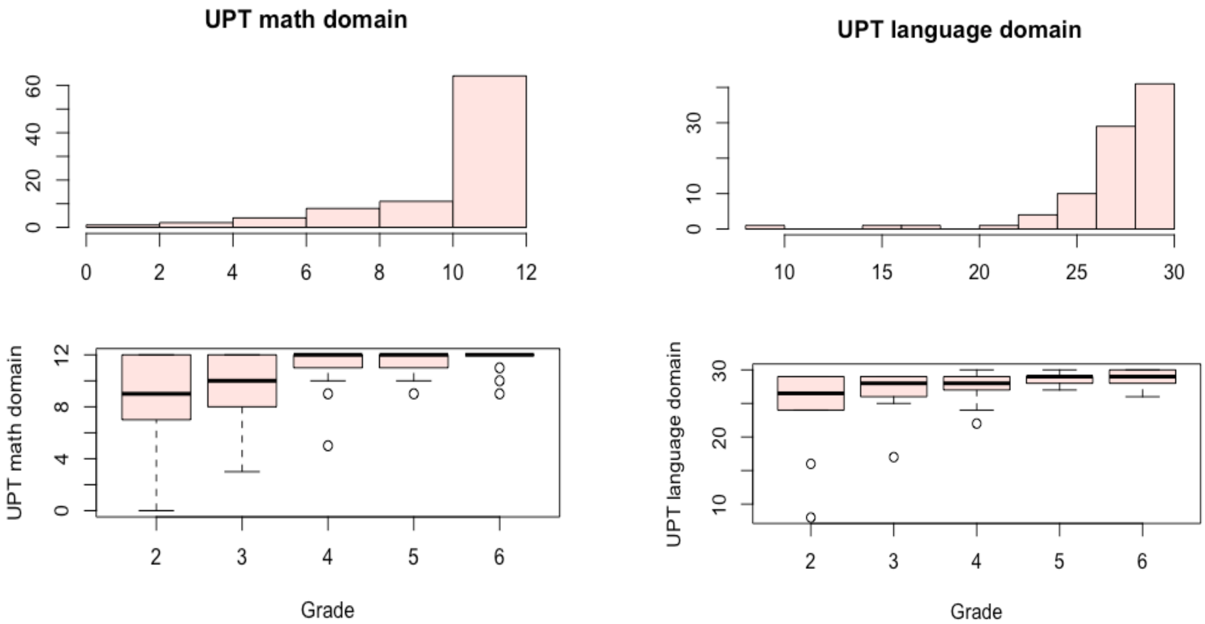


Figure 7. Distribution of correct items on the UPT across the entire sample and by grade, separated by domain.

Additionally, an analysis of outliers on the UPT was conducted. By visually inspecting the histograms and boxplots of the UPT's correct and complete scores, it was noticed that there may be some values on the lower end of the scale that could be considered outliers. Outliers were defined as observations that were outside 1.5 * IQR, where IQR (i.e., Inter Quartile Range) is the difference between the 75th and 25th quartiles. For the UPT correct scores, four observations were deemed outliers ($\text{UPT}_{\text{totalcorrect}}$=8, $\text{UPT}_{\text{totalcorrect}}$= 20, $\text{UPT}_{\text{totalcorrect}}$= 23, and $\text{UPT}_{\text{totalcorrect}}$= 24). To conduct further analyses, these outliers' scores were replaced with the nearest score in the distribution, which was a score of 32 in this case. For the UPT complete scores, six observations were deemed outliers ($\text{UPT}_{\text{totalcomplete}}$=12, $\text{UPT}_{\text{totalcomplete}}$=23, $\text{UPT}_{\text{totalcomplete}}$=27, $\text{UPT}_{\text{totalcomplete}}$=34, $\text{UPT}_{\text{totalcomplete}}$=34, and $\text{UPT}_{\text{totalcomplete}}$=35. To conduct further analyses, these outliers' scores were replaced with the nearest score in the distribution, which was a score of 37 in this case.

Academic performance, emotional and behavioural problems, and executive functioning. As predicted, raw scores of participants' academic performance showed similar developmental trends to the UPT, such that their performance increased by grade. As such, there was a significant effect of grade on performance in math calculation [WJTA; $\chi^2(4)=52.33$, $p<.001$], math fluency [WJTA; $\chi^2(4)=41.87$, $p<.001$], phonemic decoding [TOWRE; $\chi^2(4)=29.24$, $p<.001$], and single word reading [TOWRE; $\chi^2(4)=42.35$, $p<.001$]. However, these same developmental trends did not emerge for parental ratings of behavioural and emotional difficulties [SDQ; $\chi^2(4)=2.08$, $p=.72$] or EF [BDEFS-CA; $\chi^2(4)=2.38$, $p=.67$].

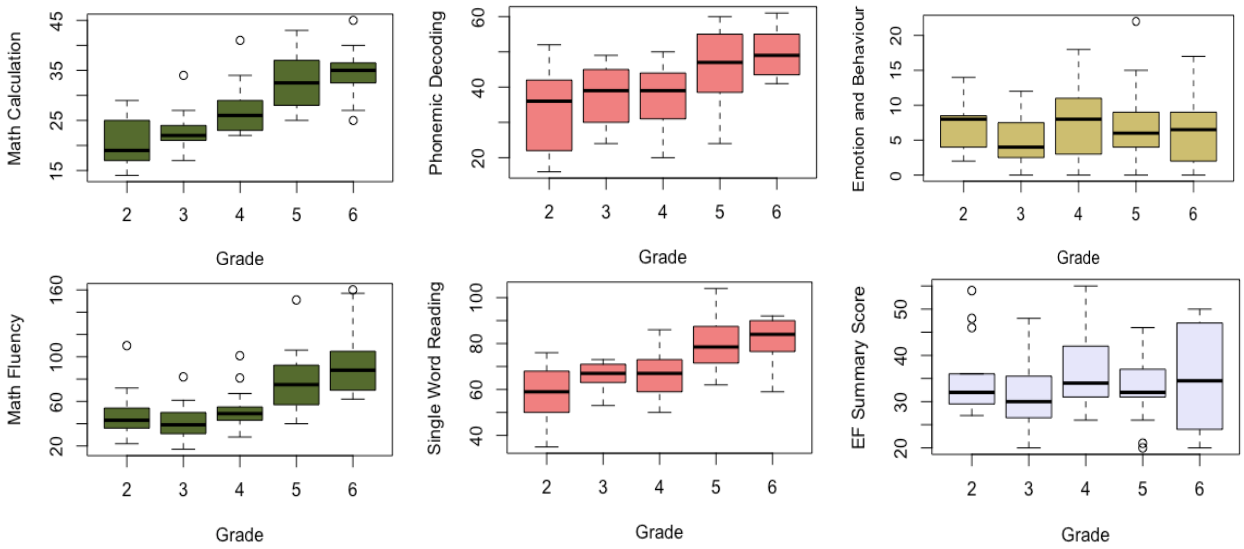


Figure 8. Distribution of academic achievement (math calculation, math fluency, phonemic decoding, single word reading), emotional and behavioural problems, and EF by grade.

Correlates of the UPT. Relationships between the UPT and academic achievement, emotional and behavioural problems, and EF were explored using Spearman correlations. All measures of academic achievement, including math calculation [WJTA; $r_s=.65$, $p<.001$], math fluency [WJTA; $r_s=.60$, $p<.001$], single word reading [TOWRE; $r_s=.47$, $p<.001$], and phonemic decoding [TOWRE; $r_s=.41$, $p<.001$] were significantly related to correct items on the UPT. However, parent-rated emotional and behavioural problems [SDQ; $r_s=-.07$, $p=.55$] and EF [BDEFS-CA; $r_s=.04$, $p=.77$] were not significantly related to correct items on the UPT (Figure 9).

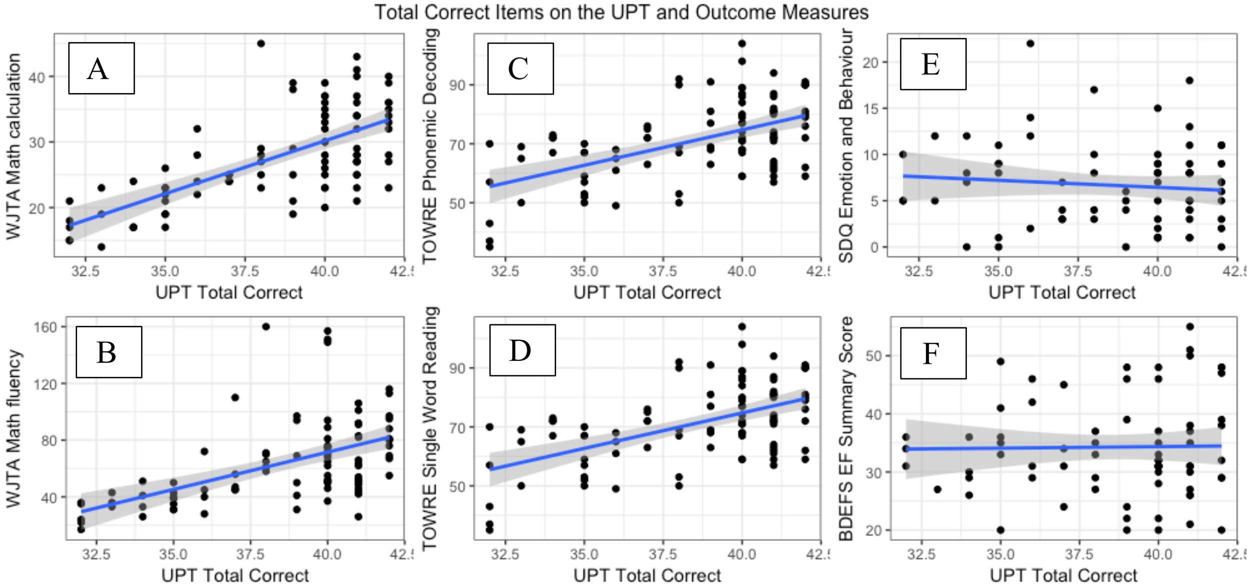


Figure 9. Scatter plots of correlations between total correct items on the UPT and various outcome measures (WJTA, TOWRE, SDQ and BDEFS-CA). The black lines represent the linear fit and the shaded areas represent the 95% confidence regions. Correct items on the UPT are shown on the x-axes and outcome measures on the y-axes, including WJTA Math Calculation (A), WJTA Math Fluency (B), TOWRE Phonemic Decoding (C), TOWRE Single Word Reading (D), SDQ Total Emotional and Behavioural Difficulties (E), and BDEFS-CA Total EF Score (F).

These relationships were further explored by separating the math and language domains of the UPT as well as the emotional and behavioural domain scores for the SDQ. As such, relationships between correct items on the total UPT, correct items in the language domain of the UPT and correct items in the math domain of the UPT and these same outcome measures were calculated using Spearman correlations. Results from these correlations are summarized in Table 4.

Table 4

Spearman correlations between correct items on the UPT and math abilities (WJTA), reading abilities (TOWRE), emotional and behavioural difficulties (SDQ), and EF (BDEFS-CA)

		UPT total correct	UPT math correct	UPT language correct
WJTA	Math calculation	.65**	.64**	.44**
	Math fluency	.60**	.63**	.41**
TOWRE	Single word reading	.55**	.41**	.39**
	Phonemic decoding	.41**	.33**	.36**
SDQ (parent)	Total difficulties	-.07	-.16	-.04
	Emotional problems	-.03	-.16	.08
	Conduct problems	-.10	-.14	-.02
	Hyperactivity	-.02	-.08	-.07
	Peer problems	-.16	-.12	-.11
	Prosocial behaviours	-.21	-.23*	-.13
BDEFS-CA (parent)	EF summary score	.04	-.01	.03

Note. * p<.05; ** p<.01. WJTA=Woodcock-Johnson Tests of Achievement, TOWRE=Test of Word Reading Efficiency, SDQ=Strengths and Difficulties Questionnaire, BDEFS-CA=Barkley Deficits in Executive Functioning Scale.

Generally, academic abilities were significantly associated with correct items on the UPT across domains, whereas parent-rated measures were not. Tests of dependent correlations using Steiger's Z test were also calculated to determine whether any of the outcome variables were significantly more related to one domain of the UPT or the other. No correlation comparisons were found to be significant, suggesting fairly consistent associations across math and language domains.

These relationships with outcome measures were further explored by examining the complete items on the UPT. Relationships between complete items on the UPT and academic achievement, emotional and behavioural problems, and EF were also explored using Spearman correlations. Parallel to findings regarding correct items on the UPT, all measures of academic achievement, including math calculation [WJTA; r_s=.40, p<.001], math fluency [WJTA; r_s=.47, p<.001], single word reading [TOWRE; r_s=.24, p=.02], and phonemic decoding [TOWRE; r_s=.27, p=.01 were significantly related to complete items on the UPT. Similarly, parent-rated emotional and behavioural problems [SDQ; r_s=-.17, p=.16] and EF [BDEFS-CA; r_s=-.01, p=.92] were not significantly related to complete items on the UPT (Figure 10).

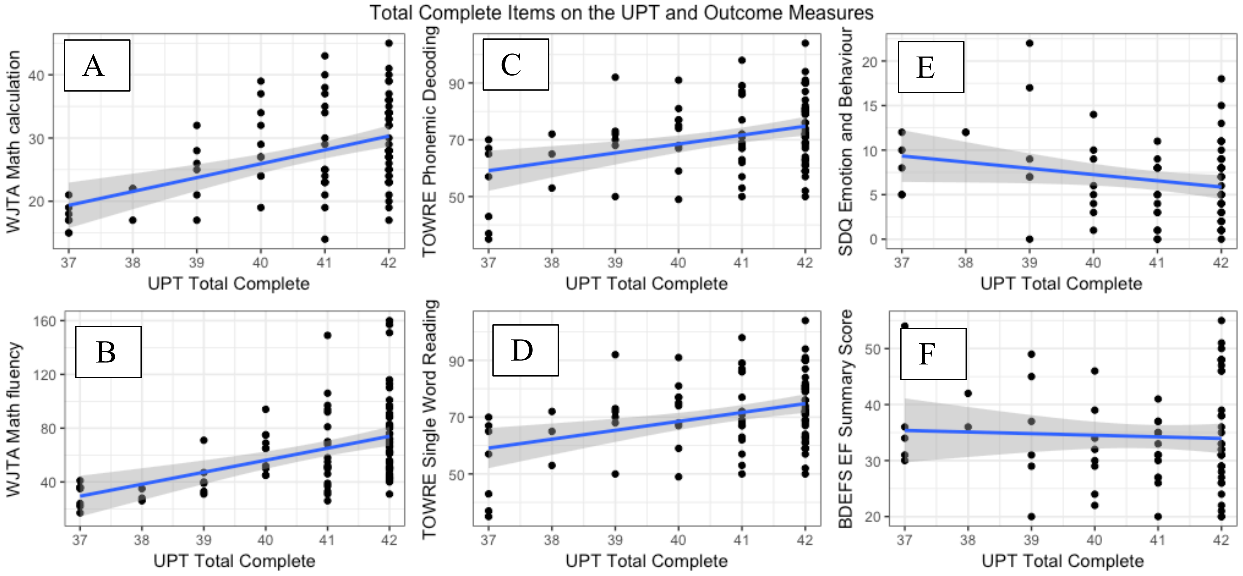


Figure 10. Scatter plots of correlations between total complete items on the UPT and various outcome measures (WJTA, TOWRE, SDQ and BDEFS-CA). The black lines represent the linear fit and the shaded areas represent the 95% confidence regions. Complete items on the UPT are shown on the x-axes and outcome measures on the y-axes, including WJTA Math Calculation (A), WJTA Math Fluency (B), TOWRE Phonemic Decoding (C), TOWRE Single Word Reading (D), SDQ Total Emotional and Behavioural Difficulties (E), and BDEFS-CA Total EF Score (F).

These relationships were further explored by separating the math and language domains of the UPT as well as the emotional and behavioural domain scores for the SDQ. As such, relationships between complete items on the total UPT, complete items in the language domain of the UPT and complete items in the math domain of the UPT and these same outcome measures were calculated using Spearman correlations. Results from these correlations are summarized in Table 5.

Table 5

Spearman correlations between complete items on the UPT and math abilities (WJTA), reading abilities (TOWRE), emotional and behavioural difficulties (SDQ), and EF (BDEFS-CA)

		UPT total complete	UPT math complete	UPT language complete
WJTA	Math calculation	.38**	.31**	.35**
	Math fluency	.47**	.38**	.39**
TOWRE	Single word reading	.24*	.17	.28**
	Phonemic decoding	.27*	.22*	.25*
SDQ (parent)	Total difficulties	-.17	-.18	-.12
	Emotional problems	-.04	-.13	.08
	Conduct problems	-.13	-.06	-.04
	Hyperactivity	-.13	-.08	-.24*
	Peer problems	-.07	-.14	.09
	Prosocial behaviours	-.14	-.14	-.07
BDEFS-CA (parent)	EF summary score	-.01	-.01	-.04

Note. * p<.05; ** p<.01. WJTA=Woodcock-Johnson Tests of Achievement, TOWRE=Test of Word Reading Efficiency, SDQ=Strengths and Difficulties Questionnaire, BDEFS-CA=Barkley Deficits in Executive Functioning Scale.

The UPT complete scores show very similar trends to the correct scores. Almost all of the measures of academic abilities (i.e. math and reading) were significantly related to both domains of the UPT, with the exception of single word reading (TOWRE), which was only significantly related to the language domain. When separating the emotional and behavioural domains on the SDQ, hyperactivity was found to be significantly related to complete items in the language domain of the UPT, [r_s=-.23, p=.04]. Tests of dependent correlations using Steiger's Z test were also calculated to determine whether any of the outcome variables were significantly more related to one domain of the UPT or the other. No correlation comparisons were found to be significant, suggesting fairly consistent associations across math and language domains.

Regression analyses. A two-stage hierarchical regression was performed to predict correct items on the UPT. Parent-rated EF (BDEFS-CA) was entered at stage one to understand the contributions of EF measures to the prediction of correct items on the UPT. All other outcome measures (i.e. academic achievement [WJTA and TOWRE] and emotional and behavioural difficulties [SDQ]) were entered in stage two, to understand the additional contributions of these variables to the prediction of correct items on the UPT-2. The first model did not significantly predict correct items on the UPT [$F(1,69)=0.05$, $p=.82$, $R^2=.001$], accounting for 0.1% of the variation in UPT correct items. The second model significantly predicted correct items on the UPT [$F(4,64)=13.18$, $p<.001$, $R^2=.42$], accounting for 42% of the variation in UPT correct items. In this model, only math and reading abilities added significantly to the prediction. Adding academic abilities (WJTA and TOWRE) and emotional and behavioural difficulties (SDQ) explained an additional 41.9% of the variation in total correct items on the UPT. Results from these analyses are summarized in Table 6.

Table 6.

Results of hierarchical regression of total correct items on the UPT.

Variable	β	SE(β)	*t*	*p*	R^2	ΔR^2
Model 1					.001	.001
Parent- rated EF (BDEFS-CA)	.01	.05	.23	.82		
Model 2					.42	.42
Parent- rated EF (BDEFS-CA)	.07	.05	1.32	.19		
Math abilities (WJTA)	2.26	.53	4.25**	<.001		
Reading abilities (TOWRE)	.99	.47	2.12*	.04		
Emotional and behavioural difficulties (SDQ)	-.001	.09	-.01	.99		

Note. * p<.05; ** p<.01. BDEFS-CA=Barkley Deficits in Executive Functioning Scale, WJTA=Woodcock-Johnson Tests of Achievement, TOWRE=Test of Word Reading Efficiency, SDQ=Strengths and Difficulties Questionnaire.

Another two-step hierarchical regression was performed to predict complete items on the UPT. Again, parent-rated EF (BDEFS-CA) was entered at stage one to understand the contributions of EF measures to the prediction of complete items on the UPT. All other outcome measures (i.e. academic achievement [WJTA and TOWRE] and emotional and behavioural difficulties [SDQ]) were entered in stage two, to understand the additional contributions of these variables to the prediction of complete items on the UPT. The first model did not significantly predict complete items on the UPT [*F*(1,69)=.32, *p*=.58, R^2=.005], accounting for 0.5% of the variation in UPT complete items. The second model significantly predicted complete items on the UPT [*F*(4,64)=5.05, *p*=.001, R^2=.19], accounting for 19% of the variation in UPT complete items. In this model, only math abilities added significantly to the prediction. Adding academic abilities (WJTA and TOWRE) and emotional and behavioural difficulties (SDQ) explained an additional 18.5% of the variation in total complete items on the UPT. Results from these analyses are summarized in Table 7.

Table 7.

Results of hierarchical regression of total complete items on the UPT.

Variable	β	SE(β)	*t*	*p*	R^2	ΔR^2
Model 1					.005	.005
Parent- rated EF (BDEFS-CA)	-.02	.03	-.56	.58		
Model 2					.19	.19
Parent- rated EF (BDEFS-CA)	.02	.04	.44	.66		
Math abilities (WJTA)	1.10	.41	2.68	.009**		
Reading abilities (TOWRE)	.35	.36	.96	.34		
Emotional and behavioural difficulties (SDQ)	-.06	.07	-.80	.42		

Note. * p<.05; ** p<.01. BDEFS-CA= Barkley Deficits in Executive Functioning Scale, WJTA=Woodcock-Johnson Tests of Achievement, TOWRE=Test of Word Reading Efficiency, SDQ=Strengths and Difficulties Questionnaire.

Additionally, a binomial logistic regression was completed to determine whether the UPT or other related outcome variables were significant predictors of status on the SDQ. Based on the SDQ total difficulties scores, participants were placed into a no-risk clinical group (SDQ total score=0-13) or an at-risk clinical group (SDQ total score=14-50) based on norms from the SDQ (Goodman, 1997). From this classification, 5 participants were placed in the at-risk clinical group, while 62 participants were placed in the no-risk clinical group. This binomial logistic model indicated that none of the outcome variables, including the UPT, significantly predicted clinical risk on the SDQ. Results from these analyses are summarized in Table 8.

Table 8.

Results of binomial logistic regression predicting SDQ clinical status.

Predictor	β	SE(β)	*z*	*p*	OR
UPT total correct items	-.05	.18	-.25	.80	.96
Math abilities (WJTA)	.49	.98	.50	.62	1.63
Reading abilities (TOWRE)	-.30	.66	-.46	.65	.74
Parent- rated EF (BDEFS-CA)	.11	.06	1.93	.06	1.12

Note. * p<.05; ** p<.01. UPT=Unstructured Performance Task, WJTA=Woodcock-Johnson Tests of Achievement, TOWRE=Test of Word Reading Efficiency, BDEFS-CA=Barkley Deficits in Executive Functioning Scale.

Updating the UPT

Based on results from Study 1a and 1b, we updated the UPT to create the UPT-2. Firstly, we aimed to address the ceiling effect that was especially apparent in typically-developing populations. In order to do so, we generated items that were of greater difficulty (e.g. more complex rhyming, math items with larger numbers, etc.) with the aim of generating more variability of scores. We ensured that only 20% of the items on this task were of greater difficulty so that the task would not become too difficult overall. Secondly, we aimed to create a double-sided version of the task with more items, totalling 50 instead of 42. With this addition, we made behavioural observations of whether the child forgot to flip the page and required a prompt to do so. Thirdly, we aimed to ensure that the task would also be appropriate for children as young as in grade 1. By lowering this age limit, we aimed to reduce the linguistic demands of the task by creating items that were more symbolic or pictorial in nature. Finally, we aimed to create a balanced number of items in the math, language and pictorial domains. As such, we included 7 easy symbolic items, 7 hard symbolic items, 9 easy math items, 9 hard math items, 9 easy language items, and 9 hard math items.

Study 2

An examination of the variables used in this study was conducted in order to ensure that assumptions were met before proceeding with further analyses. Firstly, histograms and Q-Q plots for all variables were visually inspected. Indices of skewness and kurtosis were then calculated for all variables. Shapiro-Wilks tests of normality were also conducted to assess the normality of the distribution of all variables. Results from these analyses are summarized in Table 9.

Table 9

Means, standard deviations, skewness and kurtosis indices, and W-statistics for all variables

	Mean (SD)	**Skewness**	**Kurtosis**	**W-statistics**
UPT-2 total correct	33.93 (12.91)	-0.71	-0.60	0.91**
UPT-2 total complete	43.43 (6.86)	-1.13	0.37	0.85**
WJTA math fluency	50.85 (29.93)	0.89	0.63	0.93**
WJTA math calculation	25.11 (8.48)	0.45	-0.33	0.97*
TOWRE phonemic decoding	32.15 (16.29)	-0.22	-1.10	0.96*
TOWRE single word reading	58.07 (22.05)	-0.77	-0.21	0.92**
SDQ total difficulties	6.51 (4.34)	0.89	0.91	0.94**
SDQ emotional problems	1.69 (1.85)	1.50	2.40	0.82**
SDQ conduct problems	1.05 (1.27)	1.38	1.80	0.79**
SDQ hyperactivity	2.88 (2.59)	0.54	-0.99	0.89**
SDQ peer problems	0.93 (1.30)	1.54	1.80	0.73**
SDQ prosocial behaviors	8.46 (1.81)	-0.95	-0.34	0.81**
BDEFS-CA	33.00 (7.87)	0.73	0.20	0.95**
Stroop (Interference)	40.07 (21.45)	0.45	1.80	0.96*
TMT (Part B minus Part A)	92.57 (77.52)	1.65	4.88	0.79**

Note. * p<.05; ** p<.01. UPT=Unstructured Performance Task, WJTA=Woodcock-Johnson Tests of Achievement, TOWRE=Test of Word Reading Efficiency, SDQ=Strengths and Difficulties Questionnaire, BDEFS-CA=Barkley Deficits in Executive Functioning Scale.

These analyses revealed that most of the variables in this study did not meet the assumption of normality. However, when looking at the primary outcome variable in this study (i.e., the UPT-2), it becomes apparent that the skewness and kurtosis indices are drastically improved compared to the original UPT. Nonetheless, the Shapiro-Wilks tests of normality indicate that the distributions are not normally distributed. As such, non-parametric analyses were used in this study, namely Kruskal-Wallis tests as well as Spearman correlations.

Performance on the UPT-2. Across the entire sample, participants answered an average of 33.93 (*SD*=12.91) items out of 50 correctly on the UPT-2. As with the original UPT, the number of correct items increased developmentally, according to grade level. As such, participants in grade one answered 10.33 (*SD*=4.38) items correctly, participants in grade two answered 26.82 (*SD*=8.06) items correctly, participants in grade three answered 35.39 (*SD*=9.96) items correctly, participants in grade four answered 37.69 (*SD*=8.39) items correctly, participants in grade five answered 39.07 (*SD*=6.75) items correctly, and participants in grade six answered 46.20 (*SD*=3.49) items correctly (Figure 11). A Kruskal-Wallis test revealed that there was a significant effect of grade on total correct items on the UPT-2 [$\chi^2(5)=60.91$, $p<.001$]. Post-hoc analyses revealed that grades one and two, one and three, one and four, one and five, one and six, two and three, two and four, two and five, two and six, three and six, four and six, and five and six differed significantly. There was no significant difference between males (*M*=33.00, *SD*=12.80) and females (*M*=35.16, *SD*=13.17) on UPT total correct scores [*W*=1262.5, *p*=.40]. Additionally, behavioural observations indicated that 3 participants (all of which were in grade 1) initially forgot to flip the page and required a prompt to do so. Cronbach's alpha for the total items correct on the UPT revealed good internal consistency (α=.96).

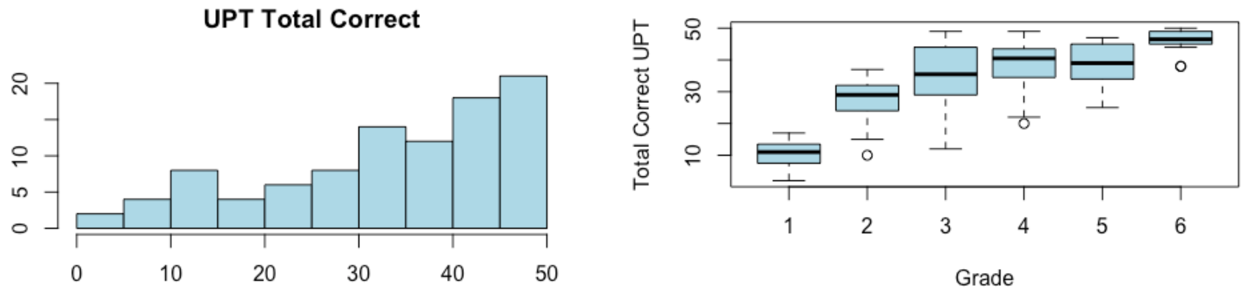


Figure 11. Distribution of correct items on the UPT-2 across the entire sample and by grade.

Across the entire sample, participants completed an average of 43.43 (*SD*=6.86) items out of 50 on the UPT. As with the original UPT, the number of items completed increased developmentally, according to grade level. As such, participants in grade one completed 33.58 (*SD*=5.02) items, participants in grade two completed 41 (*SD*=5.59) items, participants in grade three completed 44.44 (*SD*=5.75) items, participants in grade four completed 44.62 (*SD*=6.65) items, participants in grade five completed 44.29 (*SD*=5.84) items, and participants in grade six completed 48.59 (*SD*=2.04) items (Figure 12). A Kruskal-Wallis test revealed that there was a significant effect of grade on total complete items on the UPT [$\chi^2(5)=47.20$, $p<.001$]. Post-hoc analyses revealed that grades one and two, one and three, one and four, one and five, one and six, two and six, three and six, four and six, and five and six differed significantly. Cronbach's alpha for the total items complete on the UPT revealed good internal consistency (α=.90).

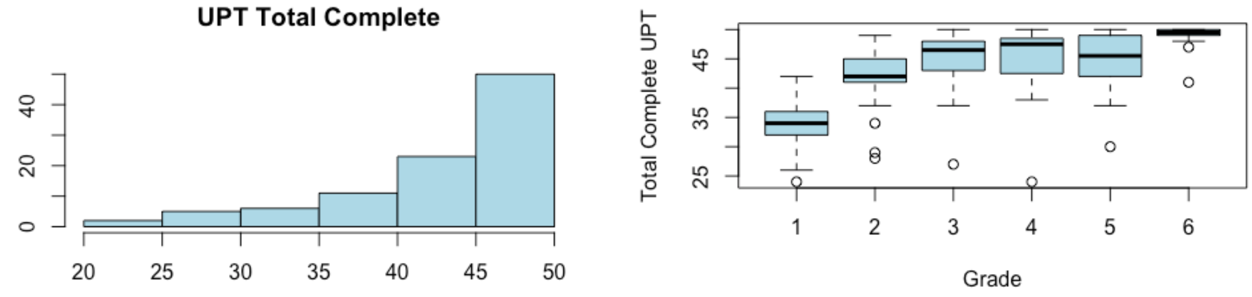
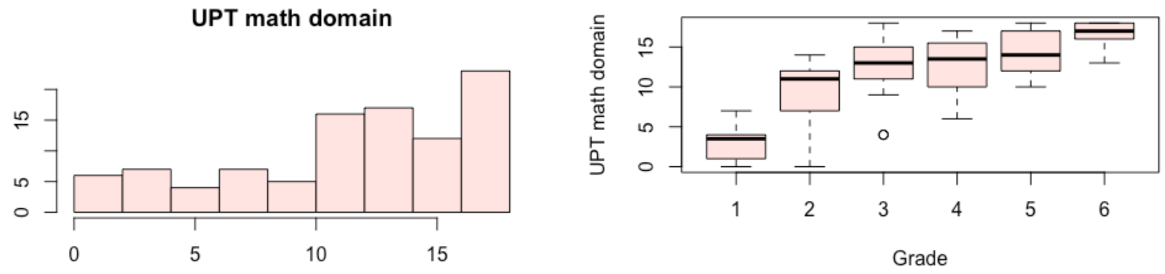


Figure 12. Distribution of complete items on the UPT-2 across the entire sample and by grade.

The UPT-2 was examined in terms of the three domains established in this updated version: math items (*n*=18), language items (*n*=18) and symbolic items (*n*=14). Across the entire sample, participants completed 66% of math items correctly, 72% of language items correctly, and 65% of symbolic items correctly. The same developmental trends identified in the total UPT-2 scores (i.e., performance increases with grade level) were also consistently found across the math, language and symbolic domains of the UPT-2 (Figure 13). Cronbach's alpha for the math domain (α=.91), language domain (α=.91), and symbolic domain (α=.88) all revealed good internal consistency.



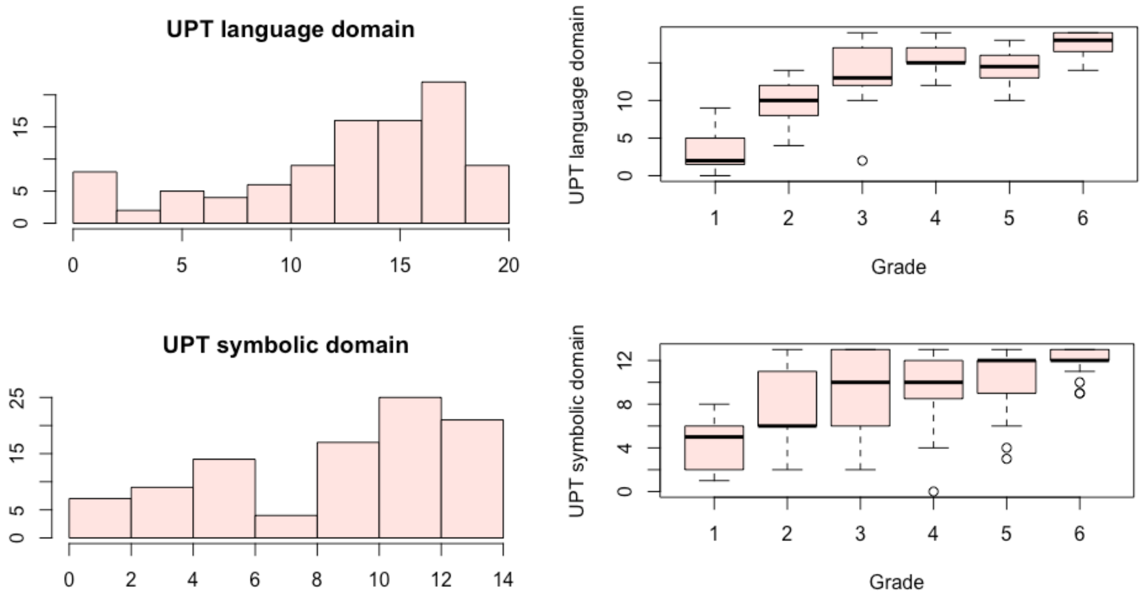


Figure 13. Distribution of correct items on the UPT-2 across the entire sample and by grade, separated by domain.

Due to the outliers that were identified on the UPT in Study 1b, we conducted an analysis of outliers in the UPT-2. Once again, outliers were defined as observations that were outside 1.5 * IQR, where IQR (i.e., Inter Quartile Range) is the difference between the 75th and 25th quartiles. Based on this analysis, there were no observations for UPT-2 correct or complete scores that were considered outliers.

Academic performance, emotional and behavioural problems, and executive functioning. As predicted, raw scores of participants' academic performance showed similar developmental trends to the UPT-2, such that their performance increased by grade. As such, there was a significant effect of grade on performance in math calculation [WJTA; $\chi^2(5)=66.57$, $p<.001$], math fluency [WJTA; $\chi^2(5)=56.35$, $p<.001$], phonemic decoding [TOWRE; $\chi^2(5)=54.34$, $p<.001$], and single word reading [TOWRE; $\chi^2(5)=61.23$, $p<.001$].

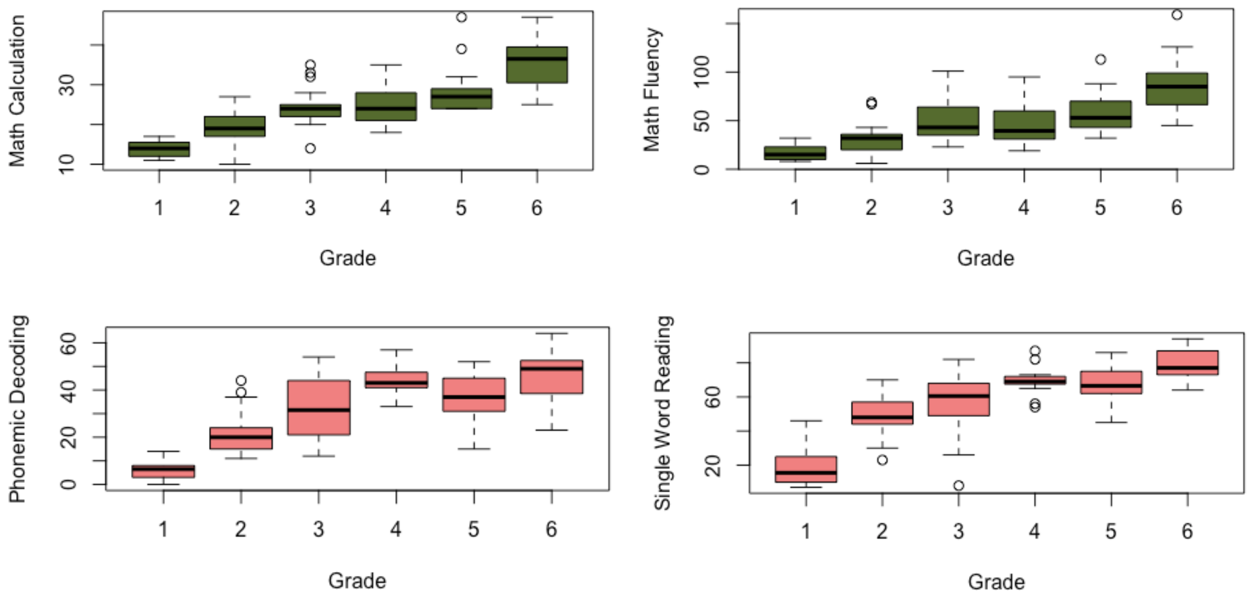
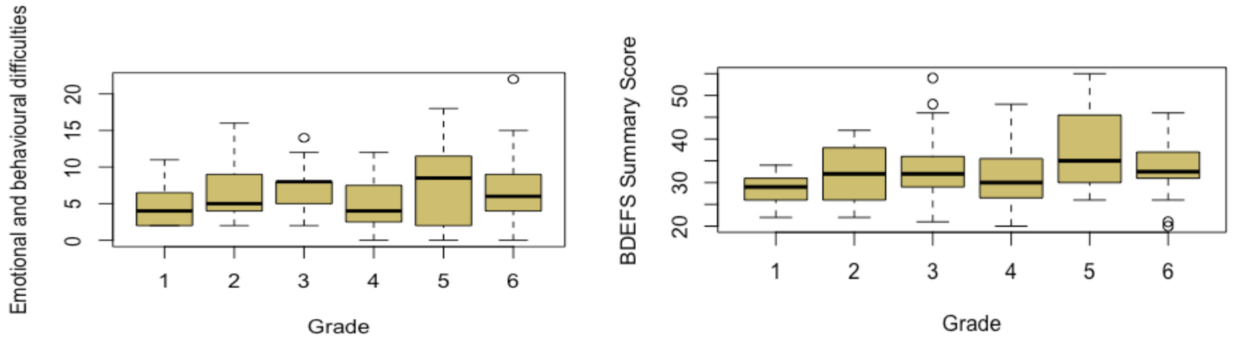


Figure 14. Distribution of academic achievement (math calculation, math fluency, phonemic decoding, single word reading) by grade.

These same developmental trends did not emerge for parental ratings of behavioural and emotional difficulties [SDQ; $\chi^2(5)=4.17$, $p=.53$] or EF [BDEFS-CA; $\chi^2(5)=6.91$, $p=.23$], suggesting that these difficulties remained fairly consistent across grade level in this sample. However, when examining EF in terms of performance-based tasks, there was a significant effect of grade on performance on the TMT [$\chi^2(5)=31.51$, $p<.001$] and the Stroop [$\chi^2(5)=25.87$, $p<.001$].



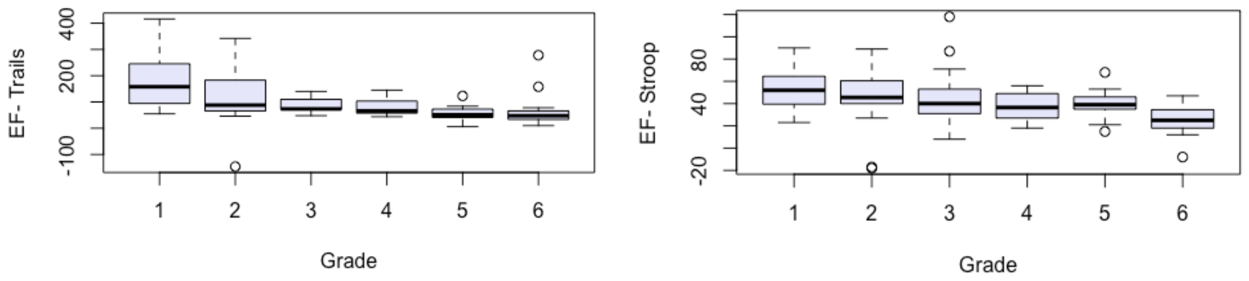


Figure 15. Distribution of parent-rated emotional and behavioural difficulties and EF, and performance-based measures of EF (Trails and Stroop) by grade.

Correlates of the UPT-2. Relationships between the correct items on the UPT-2 and academic achievement, emotional and behavioural problems, and EF were explored using Spearman correlations. As with the original UPT, all measures of academic achievement, including math calculation [WJTA; r_s=.79, p<.001], math fluency [WJTA; r_s=.74, p<.001], single word reading [TOWRE; r_s=.78, p<.001], and phonemic decoding [TOWRE; r_s=.68, p<.001] were significantly related to total correct items on the UPT-2.

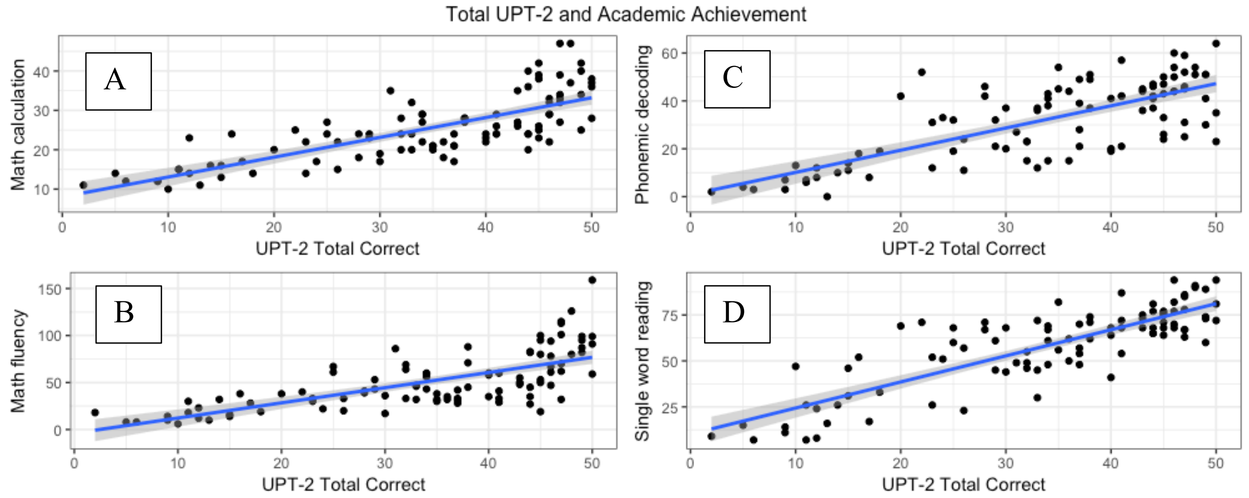


Figure 16. Scatter plots of correlations between total correct items on the UPT-2 and measures of academic achievement (WJTA and TOWRE). The black lines represent the linear fit and the shaded areas represent the 95% confidence regions. Total correct items on the UPT-2 are shown on the x-axes and outcome measures on the y-axes, including WJTA Math Calculation (A), WJTA Math Fluency (B), TOWRE Phonemic Decoding (C), and TOWRE Single Word Reading (D).

Overall parent-rated emotional and behavioural difficulties [SDQ; r_s=.01, p=.96] were not significantly related to total correct items on the UPT-2. (Figure 17).

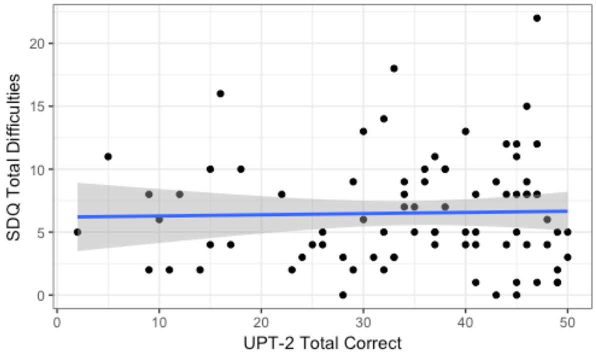


Figure 17. Scatter plot of correlation between total correct items on the UPT-2 and emotional and behavioural difficulties (SDQ). The black lines represent the linear fit and the shaded areas represent the 95% confidence regions. Total correct items on the UPT-2 are shown on the x-axis and Total Emotional and Behavioural Problems on the y-axis.

Parent-rated EF was not significantly related to total correct items on the UPT-2 [BDEFS-CA; r_s=.06, p=.57]. However, the performance-based measures of EF were both significantly related to total correct items on the UPT-2 [Trails, r_s=-.55, p<.001; Stroop, r_s=-.49, p<.001].

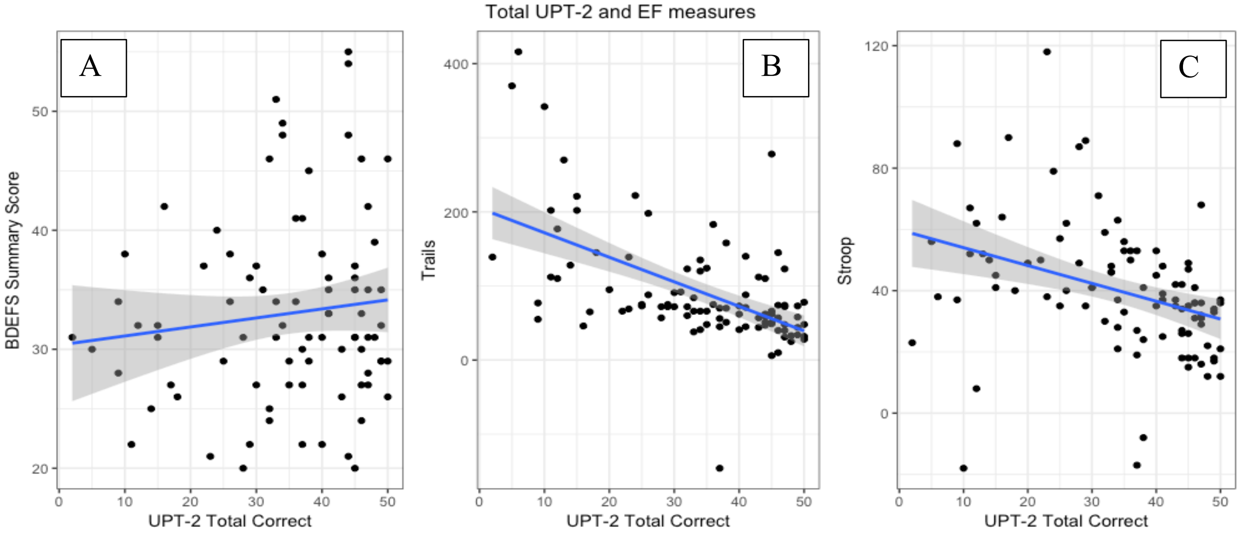


Figure 18. Scatter plots of correlations between total correct items on the UPT-2 and measures EF (BDEFS-CA, Trails, and Stroop). The black lines represent the linear fit and the shaded areas represent the 95% confidence regions. Total correct items on the UPT-2 are shown on the x-axes and outcome measures on the y-axes, including BDEFS-CA (A), Trails (B), and Stroop (C).

These relationships were further explored by separating the domains of the UPT-2 as well as the emotional and behavioural domain scores for the SDQ. As such, relationships between the total correct items on the UPT-2, correct items in the math domain of the UPT-2, correct items in the language domain of the UPT-2 and correct items in the symbolic domain of the UPT-2 and these same outcome measures were calculated using Spearman correlations. Results from these correlations are summarized in Table 10.

Table 10

Spearman correlations between correct items on the UPT-2 and math abilities (WJTA), reading abilities (TOWRE), emotional and behavioural difficulties (SDQ), parent-rated EF (BDEFS-CA), and performance-based tasks of EF (Trails and Stroop).

		UPT total correct	UPT math correct	UPT language correct	UPT symbolic correct
WJTA	Math calculation	.79**	.85**	.72**	.55**
	Math fluency	.74**	.79**	.67**	.53**
TOWRE	Phonemic decoding	.68**	.65**	.71**	.47**
	Single word reading	.78**	.75**	.80**	.54**
SDQ (parent)	Total difficulties	.01	-.04	.01	-.001
	Emotional problems	.02	.03	.05	-.01
	Conduct problems	-.04	.01	-.06	-.07
	Hyperactivity	.03	-.01	.01	.04
	Peer problems	-.03	-.07	.03	-.03
	Prosocial behaviours	.23*	.21*	.17	.28*
BDEFS-CA (parent)	EF summary score	.06	.06	.06	.06
Trails	Total score	-.55**	-.59**	-.49**	-.44**
Stroop	Total score	-.49**	-.51**	-.43**	-.39**

Note. * p<.05; ** p<.01. WJTA=Woodcock-Johnson Tests of Achievement, TOWRE=Test of Word Reading Efficiency, SDQ=Strengths and Difficulties Questionnaire, BDEFS-CA=Barkley Deficits in Executive Functioning Scale.

The UPT-2 correct scores show very similar trends to the original UPT correct scores in Study 1b. All of the measures of academic abilities (i.e. math and reading) were significantly and strongly related to all domains of the UPT-2. When separating the emotional and behavioural domains on the SDQ, prosocial behaviour was significantly related to total correct items on the UPT-2 [r_s=.23, p=.03], as well as the math [r_s=.21, p=.05] and symbolic domains [r_s=.28, p=.01] of the UPT-2. Tests of dependent correlations using Steiger's Z test were also calculated to determine whether any of the outcome variables were significantly more related to one domain of the UPT-2 or the other. No correlation comparisons were found to be significant, suggesting fairly consistent associations across the math, language and symbolic domains.

These relationships to outcome measures were further explored by examining total completed items on the UPT-2. Relationships between the complete items on the UPT-2 and academic achievement, emotional and behavioural problems, and EF were also explored using Spearman correlations. In parallel to correct items on the UPT-2, all measures of academic achievement, including math calculation [WJTA; r_s=.65, p<.001], math fluency [WJTA; r_s=.62, p<.001], single word reading [TOWRE; r_s=.68, p<.001], and phonemic decoding [TOWRE; r_s=.56, p<.001] were significantly related to total complete items on the UPT-2 (Figure 19).

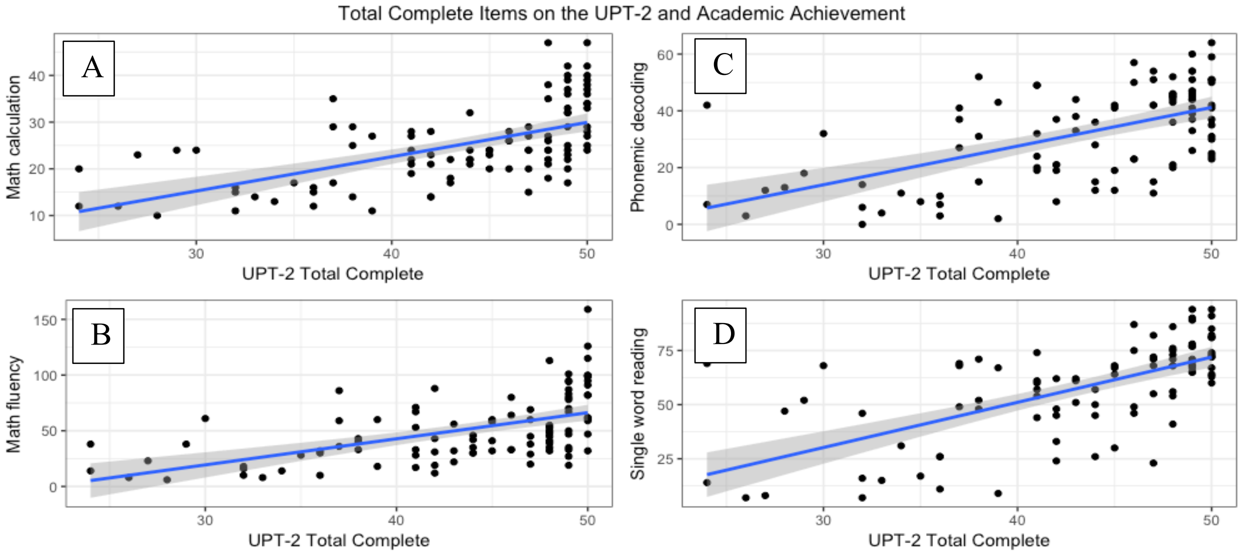


Figure 19. Scatter plots of correlations between total complete items on the UPT-2 and measures of academic achievement (WJTA and TOWRE). The black lines represent the linear fit and the shaded areas represent the 95% confidence regions. Total complete items on the UPT-2 are shown on the x-axes and outcome measures on the y-axes, including WJTA Math Calculation (A), WJTA Math Fluency (B), TOWRE Phonemic Decoding (C), and TOWRE Single Word Reading (D).

Once again, overall parent-rated emotional and behavioural difficulties [SDQ; r_s=-.05, *p*=.65] were not significantly related to total complete items on the UPT-2. (Figure 20).

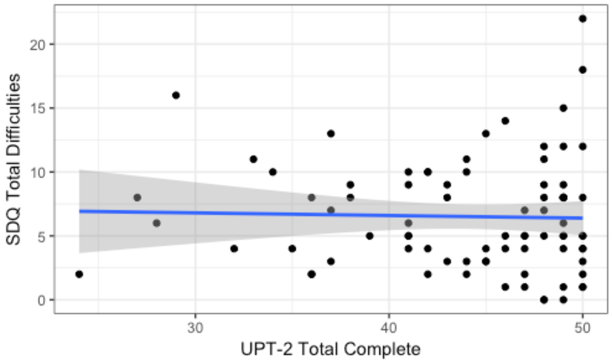


Figure 20. Scatter plot of correlation between total complete items on the UPT-2 and emotional and behavioural difficulties (SDQ). The black lines represent the linear fit and the shaded areas represent the 95% confidence regions. Total complete items on the UPT-2 are shown on the x-axis and Total Emotional and Behavioural problem on the y-axis.

As with correct items on the UPT-2, parent-rated EF was not significantly related to total complete items on the UPT-2 [BDEFS-CA; r_s=.001, *p*=.99]. However, the performance-based measures of EF were both significantly related to total correct items on the UPT-2 [Trails, *r*(95)=-.46, *p*<.001; Stroop, *r*(94)=-.36, *p*<.001] (Figure 21).

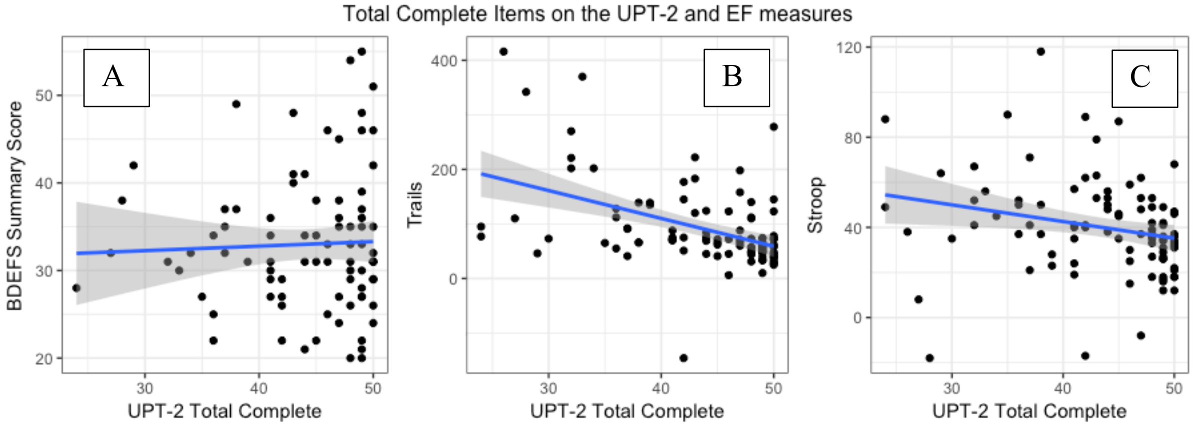


Figure 21. Scatter plots of correlations between total complete items on the UPT-2 and measures EF (BDEFS-CA, Trails, and Stroop). The black lines represent the linear fit and the shaded areas represent the 95% confidence regions. Total complete items on the UPT-2 are shown on the x-axes and outcome measures on the y-axes, including BDEFS-CA (A), Trails (B), and Stroop (C).

These relationships were further explored by separating the math, language and symbolic domains of the UPT-2 as well as the emotional and behavioural domain scores for the SDQ. As such, relationships between the total complete items on the UPT-2, complete items in the math domain of the UPT-2, complete items in the language domain of the UPT-2 and complete items in the symbolic domain of the UPT-2 and these same outcome measures were calculated using Spearman correlations. Results from these correlations are summarized in Table 11.

Table 11

Spearman correlations between complete items on the UPT-2 and math abilities (WJTA), reading abilities (TOWRE), emotional and behavioural difficulties (SDQ), parent-rated EF (BDEFS-CA), and performance-based tasks of EF (Trails and Stroop).

		UPT total complete	UPT math complete	UPT language complete	UPT symbolic complete
WJTA	Math calculation	.65**	.53**	.61**	.37**
	Math fluency	.62**	.52**	.56**	.35**
TOWRE	Phonemic decoding	.56**	.40**	.62**	.31**
	Single word reading	.68**	.49**	.73**	.34**
SDQ (parent)	Total difficulties	-.05	-.09	-.01	-.07
	Emotional problems	-.04	-.03	-.03	-.03
	Conduct problems	-.04	-.01	-.05	-.05
	Hyperactivity	.05	.01	.06	.004
	Peer problems	-.08	-.13	-.06	-.03
	Prosocial behaviours	.25*	.22*	.21*	.25*
BDEFS-CA (parent)	EF summary score	.001	-.10	.07	-.13
Trails	Total score	-.46**	-.45**	-.43**	-.27**
Stroop	Total score	-.36**	-.33**	-.37**	-.21*

Note. * p<.05; ** p<.01. WJTA=Woodcock-Johnson Tests of Achievement, TOWRE=Test of Word Reading Efficiency, SDQ=Strengths and Difficulties Questionnaire, BDEFS-CA=Barkley Deficits in Executive Functioning Scale.

Regression analyses. A two-stage hierarchical regression was performed to predict correct items on the UPT-2. The measures of EF (i.e. parent-rated EF [BDEFS-CA] and performance-based measures of EF [TMT and Stroop]) were entered at stage one to understand the contributions of EF measures to the prediction of correct items on the UPT-2. All other outcome measures (i.e. academic achievement [WJTA and TOWRE] and emotional and behavioural difficulties [SDQ]) were entered in stage two, to understand the additional contributions of these variables to the prediction of correct items on the UPT-2. The first model significantly predicted correct items on the UPT-2 [$F(2,80)=13.45$, $p<.001$, $R^2=.25$], accounting for 25% of the variation in UPT-2 correct items. In this model, only the performance-based EF tasks added significantly to the prediction. The second model significantly predicted correct items on the UPT-2 [$F(5,74)=38.75$, $p<.001$, $R^2=.70$], accounting for 70% of the variation in UPT-2 correct items. In this model, only math and reading abilities added significantly to the prediction. Adding academic abilities (WJTA and TOWRE) and emotional and behavioural difficulties (SDQ) explained an additional 45% of the variation in total correct items on the UPT. Results from these analyses are summarized in Table 12.

Table 12.

Results of hierarchical regression of total correct items on the UPT-2.

Variable	β	SE(β)	*t*	*p*	*R*²	Δ*R*²
<u>Model 1</u>					.25	.25
Parent-rated EF (BDEFS-CA)	.17	.15	1.22	.25		
Performance-based EF (TMT and Stroop)	-8.49	1.68	-5.05**	<.001		
<u>Model 2</u>					.70	.45
Parent-rated EF (BDEFS-CA)	.07	.12	.55	.59		
Performance-based EF (TMT and Stroop)	-1.66	1.20	-1.39	.17		
Math abilities (WJTA)	5.19	1.17	4.45**	<.001		
Reading abilities (TOWRE)	6.62	1.16	5.73**	<.001		
Emotional and behavioural difficulties (SDQ)	.09	.21	.43	.67		

Note. * p<.05; ** p<.01. BDEFS-CA=Barkley Deficits in Executive Functioning Scale, TMT=Trail-Making Test, WJTA=Woodcock-Johnson Tests of Achievement, TOWRE=Test of Word Reading Efficiency, SDQ=Strengths and Difficulties Questionnaire.

Another two-step hierarchical regression was performed to predict complete items on the UPT-2. Again, the measures of EF (i.e. parent-rated EF [BDEFS-CA] and performance-based measures of EF [TMT and Stroop]) were entered at stage one to understand the contributions of EF measures to the prediction of complete items on the UPT-2. All other outcome measures (i.e.

academic achievement [WJTA and TOWRE] and emotional and behavioural difficulties [SDQ]) were entered in stage two, to understand the additional contributions of these variables to the prediction of complete items on the UPT-2. The first model significantly predicted complete items on the UPT-2 [$F(2,80)=5.48$, $p=.006$, $R^2=.12$], accounting for 12% of the variation in UPT-2 complete items. In this model, only the performance-based EF tasks added significantly to the prediction. The second model also significantly predicted complete items on the UPT-2 [$F(5,74)=13.50$, $p<.001$, $R^2=.44$], accounting for 44% of the variation in UPT-2 complete items. In this model, only math and reading abilities added significantly to the prediction. Adding academic abilities (WJTA and TOWRE) and emotional and behavioural difficulties (SDQ) explained an additional 32% of the variation in total complete items on the UPT. Results from these analyses are summarized in Table 13.

Table 13.

Results of hierarchical regression of total complete items on the UPT-2.

Variable	β	SE(β)	*t*	*p*	R^2	ΔR^2
Model 1					.12	.12
Parent- rated EF (BDEFS-CA)	0.04	0.08	.56	.65		
Performance-based EF (TMT and Stroop)	-2.99	.91	-2.38**	.002		
Model 2					.44	.32
Parent- rated EF (BDEFS-CA)	-.03	.08	-.30	.77		
Performance-based EF (TMT and Stroop)	-.003	.83	-.004	.99		
Math abilities (WJTA)	1.63	.81	2.02*	.05		
Reading abilities (TOWRE)	3.44	.80	4.31**	<.001		
Emotional and behavioural difficulties (SDQ)	.02	.15	.15	.88		

Note. * p<.05; ** p<.01. BDEFS-CA=Barkley Deficits in Executive Functioning Scale, TMT=Trail-Making Test, WJTA=Woodcock-Johnson Tests of Achievement, TOWRE=Test of Word Reading Efficiency, SDQ=Strengths and Difficulties Questionnaire.

Additionally, a binomial logistic regression was completed to determine whether the UPT or other related outcome variables were significant predictors of status on the SDQ. Based on the

SDQ total difficulties scores, participants were placed into a no-risk clinical group (SDQ total score=0-13) or an at-risk clinical group (SDQ total score=14-50) based on norms from the SDQ (Goodman, 1997). From these classifications, 74 participants were classified as being at no clinical risk, while 5 participants fell in the at-risk clinical group. This binomial logistic model indicated that only parent-rated EF (BDEFS-CA) significantly predicted clinical risk on the SDQ. Results from these analyses are summarized in Table 14.

Table 14.

Results of binomial logistic regression predicting SDQ clinical status.

Predictor	β	SE(β)	*z*	*p*	OR
UPT total correct items	-.12	.10	-1.19	.23	.88
Math abilities (WJTA)	2.03	1.11	1.92	.06	7.60
Reading abilities (TOWRE)	-.44	1.12	1.82	.69	.64
Parent- rated EF (BDEFS-CA)	.24	.09	2.69	.007**	1.28
Performance-based EF (Stroop and TMT)	-1.19	1.22	-.97	.33	.31

Note. * p<.05; ** p<.0. SDQ=Strengths and Difficulties Questionnaire, WJTA=Woodcock-Johnson Tests of Achievement, BDEFS-CA=Barkley Deficits in Executive Functioning Scale, TMT=Trail-Making Test.

Discussion

The purpose of this project was to better understand key psychometric properties and correlates of the UPT in a non-clinical sample, and to subsequently update and pilot a new version of the UPT (i.e. UPT-2) to improve this novel measure. In order to do this, three related studies were undertaken. These studies had three primary objectives. First, we aimed to examine the psychometric characteristics and task structure of the UPT. Second, we aimed to examine associations between the UPT and relevant outcome measures (i.e. measures of EF, academic abilities, etc.). Third, we aimed to develop and pilot an updated version of the UPT (i.e. UPT-2) and understand core psychometric properties and correlates of this updated task.

In the first study (i.e. Study 1a), we further assessed the psychometric properties and task structure of the original UPT in a sample of children with and without ADHD. This revealed that the UPT generally had good psychometric properties and was significantly related to traditional measures of EF (Stroop, Trails, and BDEFS-CA) and parent-rated clinical domains of behaviour (CBCL). The second study (i.e. Study 1b) aimed to administer the original UPT to a new community sample of children to examine some previously unexplored correlates of the UPT (e.g. academic abilities). The UPT demonstrated comparable psychometric properties and task structure in this new community sample of children. It was also significantly related to academic abilities as assessed by reading (TOWRE) and mathematics skills (WJTA). However, the UPT was not significantly related to parent-rated EF (BDEFS-CA) or emotional and behavioural difficulties (SDQ). Based on results from studies 1a and 1b, the original UPT was updated in order to create the UPT-2. As such, the third study (i.e. Study 2) aimed to pilot the UPT-2 in a community sample of children. The UPT-2 generated greater variability in scores and improved psychometric properties in this community sample. It was also significantly related to academic abilities (TOWRE and WJTA) as well as traditional performance-based measures of EF (Trails and Stroop). However, the UPT-2 was not significantly related to parent-rated EF (BDEFS-CA) or emotional and behavioural difficulties (SDQ).

Psychometric characteristics and task structure of the UPT

The psychometric characteristics of the original UPT were assessed in Studies 1a and 1b. Specifically, in Study 1a the UPT demonstrated good internal consistency (α=.92) and split-half reliability (ρ=.90). However, it was found that internal consistency was lower for children in the control group (α=.73) than those with ADHD (α=.93). In study 1b, the UPT demonstrated good internal consistency in a community sample of children (α=.88). These indices generally demonstrate good psychometric properties of the total UPT, though there is some concern that it may not be as reliable when applied to non-clinical samples, which was an important issue to address in subsequent studies.

The issue of a potential ceiling effect on the UPT had been raised by Ledochowski et al. (2019), and we aimed to further investigate this issue in our studies. In study 1a, this ceiling effect was particularly present in the non-clinical group of children, such that the scores on the UPT had a very restricted range with many children performing near-perfectly. This was replicated in Study 1b, with this community sample of children also achieving near-perfect scores (*M*=40.21 out of 42 items), suggesting the presence of a ceiling effect. In fact, this ceiling effect that is detected in non-clinical groups of children may be in part responsible for the weaker psychometric properties of the UPT in non-clinical groups compared to the clinical group in Study 1a. Additionally, by examining the shape of the distribution of the UPT scores (i.e., skewness, kurtosis, and tests of normality), this issue was further highlighted by demonstrating that there were serious concerns in this domain. As such, the possibility of running parametric analyses on continuous data becomes impossible, which provided further motivation to address this psychometric issue. With the replication of this ceiling effect in Study 1b, we aimed to create more variability in scores with the UPT-2 in order to address this effect.

Finally, we also examined the task structure of the UPT. In Study 1a we split the UPT into two domains based on their content: math domain (*n*=12) and language domain (*n*=30). In Studies 1a and 1b, participants did not perform significantly differently in these domains as predicted. In study 1a, the language and math domains both demonstrated good internal consistency (α=.91; α=.81). However, the internal consistency of the math domain diminished considerably when looking solely at the control group (α=.40). When looking at an entirely non-clinical sample in Study 1b, internal consistency was good for both the language and math domains (α=.81; α=.84). Due to some of these mixed results regarding the math domain in particular, even proportions of math and language items were included in the UPT-2.

Correlates of the UPT

In Study 1a, the total UPT (correct items and complete items) as well as both domains (math and language) of the UPT were significantly correlated with EF tasks and EF ratings in the total sample (children with ADHD and a control group). The UPT was significantly correlated with EF task performance (i.e. Stroop, TMT and BDEFS-CA), as well as parent-rated symptoms of ADHD. This suggests that the UPT may be a useful measure of EF that may also be relevant to ADHD symptomology. Additionally, the math domain of the UPT displayed unique relationships to clinical domains of depression and anxiety, which was a novel finding for this measure.

In Study 1b, we also administered measures of academic achievement (i.e. math and reading abilities) to better understand the contributions of these core skills in the successful completion of the UPT in a non-clinical sample. Our findings demonstrated that math and reading abilities were significantly related to accuracy and completion of items on the UPT, across the math and language domains. However, contrary to our prediction, parent-rated EF and parent-rated emotional and behavioural difficulties were not significantly related to performance on the UPT in this community sample. Further analyses also demonstrated that only math and reading abilities significantly predicted scores on the UPT, and that the UPT did not predict participants' clinical status on the SDQ.

One possible explanation for the lack of relationship between the UPT (accuracy and completion scores) and these behavioural rating scales may be the ceiling effect of the UPT that was detected in this sample. However, a strong relationship was still established between the UPT and academic abilities in this sample, despite this ceiling effect on the UPT. Another explanation for this lack of relationship may be due to a restricted range on the behavioural rating scales used in this community sample. Specifically, raw scores on the BDEFS-CA ranged from 20 to 55, with most participants' scores ranging from 20 to 40. However, the possible range of scores for this measure is 20 to 80. According to Barkley (2012), scores from 20 to 40 range from about the 1st to 80th percentile, and scores from 40 to 55 range from about the 80th to 93rd percentile. As such, there is in fact a fairly restricted range of scores on the BDEFS-CA in this sample, with relatively few participants demonstrating important difficulties in EF. A similar trend appeared on the SDQ in this sample. In this sample, raw scores on the SDQ ranged from 0 to 22, with most participants' score ranging from 0 to 10. The possible range of scores for this measure is 0 to 40. From Goodman's (1997) classification of clinical risk from this measure, only 5 participants in our sample were placed in the at-risk clinical group. As such, it appears as though there are relatively few participants across this non-clinical group who may have important impairments in parent-rated EF or behavioural and emotional difficulties. These restricted ranges of scores on parent behavioural ratings may at least partly explain the lack of correlation between the UPT and parent EF ratings. Nonetheless, we aimed to address the ceiling effect of the UPT in the UPT-2 in order to better understand this issue in a non-clinical sample.

Creation and pilot of the UPT-2

Based on these findings in Studies 1a and 1b, the original UPT was updated in order to create and pilot the UPT-2. The principal aim of creating the UPT-2 was to address the ceiling effect that became apparent throughout Studies 1a and 1b. As such, the following changes were made to the UPT: (1) ensuring that 20% of the items on the UPT-2 were of greater difficulty, (2) creating a double-sided version of the task with a total of 50 items, (3) reducing linguistic demands of the task to administer it to children as young as grade one so that the instrument could be used with students from grades one to six, and (4) creating a balanced number of items in the math, language and pictorial domains.

Several methodological issues were raised when creating an updated version of the UPT. For example, when attempting to address the ceiling effect, we chose to increase the difficulty of only 20% of the items. However, increasing the difficulty of items in the UPT is an interesting issue, as the goal of the task is to be easy in terms of its content. If the items became too difficult overall, then there would be a danger that the UPT is assessing aptitude or skill of difficult mathematical or reading problems, rather than assessing an underlying construct of the task as a whole. Another possibility that we considered in order to somewhat increase the difficulty of the task was to increase the total number of items drastically. Although we did ultimately decide to increase the number of items from 42 to 50, we did not consider this change to be drastic. From a practical standpoint, there is also a potential danger in increasing the number of items substantially. Currently, the UPT/UPT-2 takes about 10-15 minutes to complete. However, if the number of items was increased substantially, it would likely surpass this time of completion. This raises questions regarding feasibility of administration in research and/or clinical setting, and whether the task will become an assessment of endurance rather than EF. These questions will continue to be raised and reflected upon throughout the continuing development of the UPT, such as creating a developmentally-appropriate version of the task for adolescents.

When the UPT-2 was administered in Study 2, there was a greater range of scores in this community-sample of children than with the original UPT. Specifically, while children correctly answered an average 96% of items correctly on the original UPT, they completed an average of 68% of items correctly on the UPT-2. Additionally, internal consistency on the UPT-2 (α=.96) was superior to that of the original UPT (α=.92). These findings suggest that the changes made to the UPT-2 created more variability in scores in terms of both accuracy and completion in a community sample, addressing the ceiling effect that was identified and replicated in the original UPT.

One of the novel features of the UPT-2 was creating a double-sided version of the task. Participants were shown that there were two sides of the task before beginning the UPT-2, and there was also a large arrow on the worksheet to remind the participant to flip the page (Appendix 2). Examiners were instructed to make a behavioural observation if the participant forgot to complete the second side of the task and required a prompt to do so. In this community sample, only three participants (all of which were in grade 1) required a prompt to complete the second side of the worksheet. As such, there was limited opportunity to conduct further analyses on this feature of the task. Nonetheless, some participants in this non-clinical sample did require this prompt, and it will be interesting to see whether this novel feature of the UPT-2 provides insight in clinical groups of children, such as those with ADHD.

The relationships of the UPT-2 with several outcome variables were also assessed in Study 2. As with the original UPT, the UPT-2 was significantly related to measures of math and reading abilities. However, contrary to predictions, the UPT-2 was not related to parent-rated EF or parent-rated behavioural and emotional difficulties. Nonetheless, the UPT-2 was significantly related to performance-based tasks of EF as predicted. When predicting scores on the UPT-2 using hierarchical regressions, only reading and math abilities predicted performance on the UPT-2 in the final model. However, it is difficult to confidently parse apart the independent contributions of EF abilities and academic abilities in performance on the UPT, as the literature shows that developing EF abilities in children are importantly associated to basic mathematical and reading abilities (van der Sluis, de Jong & van der Leij, 2007; Cragg, Keeble, Richardson, Roome & Gilmore, 2017). Additionally, the UPT-2 did not significantly predict clinical risk on the SDQ. Parent-rated EF did predict clinical risk on the SDQ, which is to be expected as both the BDEFS-CA and the SDQ are parent-rated measures of behaviour.

Given that the UPT-2 is intended to act as a measure of assessing EF in children, we expected that the BDEFS-CA would be significantly related to the UPT-2 and that the EF measures would predict scores on the UPT-2, as was found by Ledochowski et al. (2019). Although greater variability of scores was achieved in the UPT-2, we still did not detect expected relationships to the BDEFS-CA and the SDQ. Parallel to Study 1b, participants' scores on these behavioural rating measures demonstrated restricted ranges, with very few parents reporting that their child experienced important difficulties in EF or emotional and behavioural abilities. Specifically, raw scores on the BDEFS-CA ranged from 20 to 55, with most participants' scores ranging from 20 to 35. However, the possible range of scores for this measure is 20 to 80. According to Barkley (2012), scores from 20 to 35 range from about the 1st to 70th percentile, and scores from 35 to 55 range from about the 71st to 93rd percentile. As such, there is a fairly restricted range of scores on the BDEFS-CA once again, with relatively few participants demonstrating important difficulties in EF. A similar trend appeared on the SDQ in this sample. In this sample, raw scores on the SDQ ranged from 0 to 22, with most participants' score ranging from 0 to 10. The possible range of scores for this measure is 0 to 40. From Goodman's (1997) classification of clinical risk from this measure, only 5 participants in our sample were placed in the at-risk clinical group. As such, it appears as though there are, again, relatively few participants across this non-clinical group who may have notable impairments in parent-rated EF or behavioural and emotional difficulties. Once again, the restricted range of scores of these behavioural rating measures in this sample may explain the lack of relationship to the UPT-2.

It is also important to note that the samples used in Study 2 was an exclusively non-clinical group, whereas Ledochoswki et al. (2019) included a group of children with ADHD in their analyses. Perhaps due to the intended easy nature of this task, the UPT does not impose the same EF demands in a non-clinical group that it does for a clinical group with known EF difficulties, such as those with ADHD. As such, it is possible that the UPT acts as a measure of impairment. For example, inability to complete this relatively easy task may be an indicator of impairment as opposed to degrees of performance on the UPT being systematically related to degrees of difficulty indicated on behaviour rating scales. That may, for example, explain why the predicted patterns of association between the UPT and behavioural rating scales were detected in a sample that includes a clinical group (Ledochowski et al., 2019), but not in an exclusively community sample. As such, it would be interesting to further examine these relationships with the UPT-2 in a clinical sample of children as well.

Additionally, there are a host of factors and issues to consider when discussing the construct validity and correlates of a new measure, particularly with measures of EF. For example, the "task-impurity problem" is frequently discussed within the context of EF measures. Burgess (1997) explains that many EF tasks incidentally measure a range of related executive processes, which likely accounts for the large measurement error in these tasks. Due to this issue, conclusions regarding EF deficits from a single EF task are often questionable due to the many sub-constructs that are embedded within this construct (Lambek et al., 2011). As such, it is important to continue considering measurement issues of the UPT, such as the task-impurity problem, in order to best understand and conceptualize correlates of this novel task. Nonetheless, this task may provide valuable insight into how a child uses self-direction abilities in unstructured situations, which is lacking in our current assessment tools of EF.

Developmental considerations

One important perspective that was maintained throughout this project was the adoption of a developmental lens to increase our understanding of the UPT throughout childhood. While

Ledochowski et al. (2019) aimed to understand performance on the UPT across clinical and non-clinical groups of children, this project extended this work by better conceptualizing performance on the UPT across development in a community sample. In both the UPT and the UPT-2, we examined and identified developmental differences on this task, such that performance generally increased with age and grade level. Interestingly, we did not identify parallel developmental trajectories of behavioural ratings of EF (BDEFS-CA) or emotional and behavioural difficulties (SDQ). It is important to consider that EF abilities gradually improve over the course of development (Best & Miller, 2011). Therefore, the developmental differences that we identified on the UPT may highlight the developmental nature of EF abilities in childhood.

However, when identifying these developmental differences on the UPT, it is also important to consider the contribution of developing academic abilities. The significant relationship between performance on the UPT and the UPT-2 and reading and math abilities may in part explain the developmental trajectory identified on this task. It is hard to further parse apart these relationships at this time to better understand the contribution of developing academic abilities and developing EFs to performance on the UPT. This issue becomes further complicated by the fact that developing EF abilities and developing academic abilities have complex interactive relationships with one another (Ribner et al., 2017). Nonetheless, this issue may also provide interesting insight in regard to further developments of this measure. If it is deemed that developing academic abilities are largely driving performance on the UPT, perhaps it may be more appropriate to create a separate version of the UPT for each grade. Rather than administering the same measure to children from grades 1 to 6, this method may prove helpful in isolating the independent contributions of EF and academic abilities to performance on the UPT.

Additionally, we extended the age range of the UPT-2, such that children as young as grade 1 (i.e. age 5-6) completed the task, compared to the original UPT which was designed for children aged 8-12. Although younger participants certainly performed more poorly on the UPT-2 than older participants, they were still generally able to complete the task in an age-appropriate manner without the presence of a floor effect. This finding affirmed that the UPT-2 appears to be a developmentally-appropriate tool for children aged 6-12. This increased age range allows for greater utility of the task, such that the UPT-2 could be used as early detection of potential EF difficulties, and could potentially be used to follow the developmental trajectory of a child's EF abilities as they age.

Importance and contribution

This project further investigated and improved the UPT as a novel performance-based measure of EF. By further studying the UPT and creating the UPT-2, we generated greater variability of scores in a community sample, consequently addressing the ceiling effect detected in the original UPT. Additionally, the UPT-2 demonstrated developmental sensitivity for children aged 6-12. Finally, we also demonstrated that the UPT-2 is reliable across domains of the task, which includes math, language, and symbolic domains.

Another contribution of this project was developing a greater understanding of the UPT and relevant data patterns in a community sample. Explicitly studying the development of core cognitive abilities, such as EFs, in non-clinical children contributes importantly to our understanding of developmental trajectories and differences in cognition. This, in turn, can help inform our conceptualization of these key abilities in children, including those who may present with developmental differences. Without this understanding of cognitive development, it may be difficult for clinicians and researchers to accurately conceptualize children with developmental difficulties. For example, ADHD is currently conceptualized as a neurodevelopmental disorder

(American Psychiatric Association, 2013). Despite this conceptualization, there are still many questions regarding the developmental patterns of ADHD, such as how symptoms vary with age. Additionally, there is a wide range of EF abilities within groups of children without clinical concerns that can help shape many important abilities in their current lives, as well as provide interesting insight for their outcomes in adulthood (Blair, 2016). As such, it is essential to have tools of EF that demonstrate developmental sensitivity for children with and without clinical concerns.

Overall, the UPT is a novel performance-based measure of EF that assesses how a child would perform in an unstructured situation, requiring them to use their self-direction abilities to perform the desired task. This task may provide valuable insight into how a child uses these self-direction abilities in unstructured situations, such as the classroom or home environment. This type of performance-based EF task is currently lacking in the clinical and experimental tools that we have. The standardized measures that we currently have at our disposal can potentially miss some of the EF difficulties that children may typically experience due to their structured nature. For example, children with ADHD can often present as quite attentive and calm in a one-on-one standardized setting, despite contrary reports regarding their daily behaviour (Mahone & Hoffman, 2007).

It is important to note that there are several existing performance-based measures of EF in the literature that have minimized the amount of structure imposed by the examiner. Despite efforts being made in these tasks to reduce the structure of the task and the involvement of the examiner, they still tend to impose some structured components, which reduces the amount of self-direction required of the child. For example, the design fluency test from the well-known Delis-Kaplan Executive Function System allows the child to create any designs that they want. However, there are rules to each of the conditions administered as well as an imposed time limit

(Delis, Kaplan & Kramer, 2001). Another example is the Zoo Map Test 1 & 2 from the BADS-C. While the Zoo Map Test 1 is an open-ended task with little structure provided, the Zoo Map Test 2 has a much more imposed structure, which reduces the child need to plan and self-direct (Emslie et al., 2003). As such, there are currently limited options available to researchers and clinicians who may wish to select an unstructured performance-based task to assess a child's use of self-direction abilities.

The construct of self-direction is importantly related to several current conceptualizations of EF in the literature. For example, Barkley's "hybrid" theory of EF postulates that there are at least six self-directed executive activities that are used to choose goals and subsequently select, enact, and sustain actions to accomplish these goals. These include self-inhibition, self-directed sensory-motor action, self-directed private speech, self-direction emotion/motivation, self-directed play and self-directed attention (Barkley, 2012). Despite playing such a fundamental role in key theories of EF, it is surprising that many performance-based tasks of EF fail to assess self-direction in children. As such, the UPT's focus on measuring self-direction in children helps to fill this important gap.

Clinical implications

Executive functioning has crucial implications for the cognitive development of children. In fact, EF abilities are shown to help foster other key skills in children such as positive peer relations, intellectual abilities and academic abilities (Arffa, 2007; Diamantopoulou et al., 2007). As such, it is essential that we have tools that can accurately and efficiently assess these abilities in children throughout their development. The UPT was designed as an ecologically valid tool of EF that more closely resembles the EF demands in the everyday lives of children. Although there were some promising results regarding the continuing development of the UPT in these studies, there remain some questions to be answered in order to determine the potential clinical utility of the UPT. For example, despite promising results of the UPT-2 in the community sample in this study, it is necessary to also administer this version of the task to clinical populations (e.g., children with ADHD) to understand the data patterns in these groups. Additionally, as EF difficulties are often conceptualized as being transdiagnostic (Snyder, Miyake & Hankin, 2015), it would be clinically useful and relevant to administer the UPT to children beyond those with ADHD.

Moving forward with this task, the UPT may be used by clinicians and researchers as a promising measure of self-directed EF abilities, with important implications for children with neurodevelopmental conditions such as ADHD. We believe that the UPT could potentially be used clinically as a screener of EF difficulties related to self-direction. Rather than serving as a diagnostic tool of EF difficulties in children, the UPT could be used as a brief indicator of how a child uses their self-direction abilities in an unstructured setting, which can complement our current assessments of EF in children.

Future directions

There are several future steps that could be undertaken in order to continuing evaluating the utility of the UPT as a clinical tool to assess EF in children. Firstly, the UPT-2 should be tested in a clinical sample of children with EF difficulties, such as those with ADHD, in order to better establish the correlates and utility of this task as a performance-based measure of EF while also controlling for the influence of reading and math abilities. Secondly, future research should concretely investigate the contribution of the element of structure in the UPT. For example, a parallel instrument could be designed with items presented in a structured grid, rather than in an unstructured and random way. From there, comparing performance of participants on the UPT as well as the parallel structured form would allow for further understanding of the role of structure in performance-based tasks of EF. Thirdly, adapting the UPT for adolescents would allow for the assessment of EF more broadly across the span of development. Additionally, this would allow for re-testing as children age, which could prove to have important clinical utility.

Limitations

There are some important limitations to note in this project. Firstly, the community samples of children recruited from JICS in Study 1b and Study 2 were generally considered to be a high-performing sample. As JICS is a private laboratory school at the University of Toronto, children in these samples tended to be of higher SES and have highly educated parents, for example. This may have limited the variability in results as well as the generalizability of the findings. As such, it would be important to replicate this work on the UPT in a community sample with more normative SES levels. Secondly, for the sake of brevity of the assessment and to limit measurement burden, there was only one measure administered for each construct. Similarly, this may limit the generalizability of the findings, as the measures selected may not have fully encompassed the constructs being studied. This may be particularly relevant to measures of EF, as EF is a multi-dimensional and complex construct. Thirdly, there were restricted ranges of scores on parent-rated behaviour scales in Studies 1b and 2. As was largely discussed earlier, this limited the generalizability of many of the conclusions that could be drawn from this study. It is particularly relevant to note that this greatly impacted the binomial logistic regressions, due to the fact that the "clinical-risk" and "no clinical-risk" groups were very unevenly balanced. As such, the conclusions drawn from these analyses in particular should be interpreted with caution and replicated in a sample with greater variability.

Conclusion

The current project examined and further developed a novel Unstructured Performance Task (UPT) to assess EF in children. Study 1a demonstrated that the UPT generally had good psychometric properties, and further elucidated the task structure of the UPT by examining the math and language domains separately. Study 1b further investigated the UPT in a community sample of children, demonstrating that the UPT was significantly related to academic abilities in children, but had limited relationships to parent-rated behaviour and EF measures. In both of these studies, it was found that the UPT demonstrated a ceiling effect, particularly in non-clinical children. From these results, the UPT was updated to create the UPT-2 and subsequently piloted in a community sample of children in Study 2. The UPT-2 was found to be significantly related to performance-based tasks of EF as well as academic abilities. Overall, results indicate that the UPT-2 may be a promising measure of EF in children, providing insight into the way that children employ their EF abilities in an unstructured setting.

References

Achenbach, T. M. & Rescorla, L. A. (2001). *The manual for the ASEBA school-age forms & profiles*. Burlington, VA: University of Vermont, Research Center for Children, Youth, and Families.

Adler, L. A., Faraone, S. V., Spencer, T. J., Berglund, P., Alperin, S. & Kessler, R. C. (2017). The structure of adult ADHD. *International Journal of Methods in Psychiatric Research, 26*(1).

Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. *Child Neuropsychology, 8*(2), 71-82.

Arffa, S. (2007). The relationship of intelligence to executive function and non-executive function measures in a sample of average, above average, and gifted youth. *Archives of Clinical Neuropsychology, 22*(8), 969-978.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Washington, DC: American Psychiatric Association.

Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions. *Psychological Bulletin, 121*, 65-94.

Barkley, R. A. (2012). *Barkley Deficits in Executive Functioning Scale – Children and Adolescents (BDEFS-CA-CA)*. Guildford Press.

Barkley, R. A. & Fischer, M. (2011). Predicting impairment in major life activities and occupational functioning in hyperactive children as adults: Self-reported executive function (EF) deficits versus EF tests. *Developmental Neuropsychology, 36*(2), 137-161.

Barkley, R. A. & Murphy, K. R. (2010). Impairment in occupational functioning and adult ADHD: The predictive utility of executive function (EF) ratings versus EF tests. *Archives of Clinical Neuropsychology, 25*(3), 157-173.

Best, J. R. & Miller, P. H. (2011). A developmental perspective on executive function. *Child Development, 81*(6), 1641-1660.

Blair, C. (2016). Developmental science and executive function. *Current Directions in Psychological Science, 25*(1), 3-7.

Bodnar, L. E., Prahme, M. C., Cutting, L. E., Denckla, M. B. & Mahone, E. M. (2007). Construct validity of parent ratings of inhibitory control. *Child Neuropsychology, 13*(4), 345-362.

Brocki, K. C. & Bohlin, G. (2004). Executive functions in children aged 6 to 13: A dimensional and developmental study. *Developmental Neuropsychology, 26*, 571-593.

Brod, G., Bunge, S. A. & Shing, Y. L. (2017). Does one year of schooling improve children's cognitive control and alter brain activation? *Psychological Science, 28*(7), 967-978.

Bunge, S. A., & Zelazo, P. D. (2006). A brain-based account of the development of rule use in childhood. *Current Directions in Psychological Science, 15*(3), 118-121.

Burgess, P. W. (1997). Theory and methodology in executive function research. In P. Rabbitt (Ed.), *Methodology of frontal and executive function* (pp. 81-116). Hove, UK: Psychology Press.

Carpendale, J. & Lewis, C. (2006). *How children develop social understanding*. Oxford, UK: Blackwell Publishing.

Cirino, P. T., Ahmed, Y., Miciak, J., Taylor, W. P., Gerst, E. H. & Barnes, M. A. (2018). A framework for executive function in the late elementary years. *Neuropsychology, 32*(2), 176-189.

Clark, C., Prior, M. & Kinsella, G. J. (2000). Do executive function deficits differentiate between adolescents with ADHD and oppositional defiant/conduct disorder? A neuropsychological study using the Six Elements Test and Hayling Sentence Completion Test. *Journal of Abnormal Child Psychology, 28*(5), 403-14.

Cragg, L., Keeble, S., Richardson, S., Roome, H. E. & Gilmore, C. (2017). Direct and indirect influences of executive functions on mathematics achievement. *Cognition, 162*, 12-26.

Davidson, M. C., Amso, D., Anderson, L. C. & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia, 44*(11), 2037-2078.

Delis, D. C., Kaplan, E. & Kramer, J. H. (2001). *Delis-Kaplan Executive Function System examiner's manual*. San Antonio, TX: NCS Pearson Inc.

Diamantopoulou, S., Rydell, A-M., Thorell, L. B., Bohlin, G. (2007). Impact of executive functioning and symptoms of attention deficit hyperactivity disorder on children's peer relations and school performance. *Developmental Neuropsychology, 32*(1), 521-542.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology, 64*, 135-68.

Emslie, H., Wilson, F. C., Burdern, V., Nimmo-Smith, I. & Wilson, B. A. (2003). *Behavioural assessment of the dysexecutive syndrome for children (BADS-C)*. London, England: Harcourt Assessment.

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

Faraone, S. V., Asherson, P., Banaschewski, T., Biederman, J., Buitelaar, J.K., Ramos-Quiroga, J.A., Rohde, L. A., Sonuga-Barke, E. J., Tannock, R. & Franke, B. (2015). Attention-deficit/hyperactivity disorder. *Nature Reviews Disease Primers, 1*, 15020.

Fisher, L., Lucas, C., Sarsfield, J. & Shaffer, D. (2006). *Interviewer manual*. Columbia University DISC Development Group.

Friedman, N. P. & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex, 86*, 186-204.

Gioia, G. A., Isquith, P. K., Guy, S. C. & Kentworthy, L. (2000). *Behavioural rating inventory of* *executive function*. Florida: Pscyhological Assessment Resources.

Goeman, J. J. & Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science, 26*(4), 584-597.

Golden, C. J. (1978). Stroop colour and word test. *Age, 15*, 90.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, 38*, 581-586.

Gray, S. A., Fettes, P., Woltering, S., Mawjee, K. & Tannock, R. (2016). Symptom manifestation and impairments in college students with ADHD. *Journal of Learning Disabilities, 49*(6), 616-630.

Gregory, R. J. (2011). *Psychological testing history, principles, and applications (6th Ed.)*. Boston: Pearson Inc.

Heaton, R. K., Chelune, G. F., Talley, J. L., Kay, G. G. & Curtis, G. (1993). *Wisconsin Card Sorting Test Manual: Revised and Expanded*. Flordia: Psychological Assessment Resources.

Hughes, C. (2002). Executive functions and development: Emerging themes. *Infant and Child Development, 11*, 201-209.

Huizinga, M., Dolan, C. V. & van der Molen, M. W. (2006). Age-related change in executive function: Developmental trends and a latent variable analysis. *Neuropsychologica, 44*(11), 2017-2036.

Jewsbury, P. A., Bowden, S. C., & Strauss, M. E. (2016). Integrating the switching, inhibition, and updating model of executive function with the Cattell-Horn-Carroll model. *Journal of Experimental Psychology: General, 145*(2), 220-245.

Kane, M. J., Brown, L. H., McVay, J. C., Silvia, P. J., Myin-Germeys, I. & Kwapil, T. R. (2007). For whom the mind wanders, and when: An experience-sampling study of working memory and executive control in daily life. *Psychological Science, 18*(7), 614-621.

Kaufman, A. S. & Kaufman, N. L. (2004). *Kaufman brief intelligence test*. John Wiley & Sons, Inc.

Lambek, R., Tannock, R., Dalsgaard, S., Trillingsgaard, A., Damm, D. & Thomsen, P. H. (2011). Executive dysfunction in school-age children with ADHD. *Journal of Attention Disorders, 15*(8), 646-655.

Ledochowski, J., Andrade, B. F. & Toplak, M. E. (2019). A novel unstructured performance-based task of executive function in children with attention-deficit/hyperactivity disorder. *Journal of Clinical Experimental Neuropsychology, 4*, 1-15.

Lehto, J. E., Juujarvi, P., Kooistra, L. & Pulkkinen, L. (2003). Dimensions of executive functioning: Evidence from children. *British Journal of Developmental Psychology, 21*, 59-80.

Mahone, E. M., Cirino, P. T., Cutting, L. E., Cerrone, P. M., Hagelthorn, K. M. Hiemenz, J. R., Singer, H. S. & Denckla, M. B. (2002). Validity of behavior rating inventory of executive function in children with ADHD and/or Tourette syndrome. *Archives of Clinical Neuropsychology, 17*(7), 643-662.

Mahone, E. M. & Hoffman. (2007). Behavior ratings of executive function among preschoolers with ADHD. *The Clinical Neuropsychologist, 21*(4), 569-586.

Martin, J. & Failows, L. (2010). Executive function: Theoretical Concerns. In Sokol, B., Muller, U., Carpendale, J., Young, A. & Iarocci, G. (Eds.), *Self- and social-regulation: Exploring the relations between social interaction, social understanding, and the development of executive functions*. doi: 10.1093/acprof:oso/9780195327694.001.0001

McAuley, T., Chen, S., Goos, L., Schachar, R. & Crosbie, J. (2010). Is the behavior rating inventory of executive function more strongly associated with measures of impairment or executive function? *Journal of the International Neuropsychology Society, 16*(3), 495-505.

McLuckie, A., Landers, A. L., Rowbotham, M., Landine, J., Schwartz, M., & Ng, D. (2018). Are parent- and teacher-reported executive function difficulties associated with parenting stress for children diagnosed with ADHD? *Journal of Attention Disorders*, 1087054718756196.

Messer, D., Bernardi, M., Botting, N., Hill, E. L., Nash, G., Leonard, H. C. & Henry, L. A. (2018). An exploration of the factor structure of executive functioning in children. *Frontiers in Psychology, 9*, 1179.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A. & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49-100.

Moriguchi, Y., & Hiraki, K. (2009). Neural origin of cognitive shifting in young children. *Proceedings of the National Academy of Sciences of the United States of America, 106*(14), 6017-6021.

Nigg, J. T. (2009). *What causes ADHD?: Understanding What Goes Wrong and Why*. New York, NY: Guilford Publications.

Reitan, R. M. (1971). Trail Making Test results from normal and brain-damaged children. *Perceptual and Motor Skills, 33*(2), 575-581.

Ribner, A. D., Willoughby, M. T., Blair, C. B. & The Family Life Project Key Investigators (2017). Executive function buffers the association between early math and later academic skills. *Frontiers in Psychology, 8*, 869.

Roberts, R. J. & Pennington, B. F. (1996). An interactive framework for examining prefrontal cognitive processes. *Developmental Neuropsychology, 12*, 105-126.

Roth, R. M., Isquith, P. K. & Gioia, G. A. (2005). *Behavior rating inventory of executive function – adult version (BRIEF-A)*. Lutz, FL: Psychological Assessment Resources.

Salthouse, T. A., Atkinson, T. M. & Berish, D. E. (2003). Executive functioning as a potential mediator of age-related cognitive decline in normal adults. *Journal of Experimental Psychology: General, 132*(4), 566-594.

Schachar, R., Gordon, L. D., Robaey, P., Chen, S., Ickowicz, A. & Barr, C. (2007). Restraint and cancellation: Multiple inhibition deficits in attention deficit hyperactivity disorder. *Journal of Abnormal Child Psychology, 35*(2), 229-238.

Schrank, F. A., McGrew, K. S., Mather, N. & Woodcock, R. W. (2014). *Woodcock-Johnson IV*. Rolling Meadows, IL: Riverside Publishing.

Shaw, P., Eckstrand, K., Sharp, W., Blumenthal, J., Lerch, J. P., Greenstein, D., Clasen, L., Evans, A., Giedd, J. & Rappaport, J. L. (2007). Attention-deficit/hyperactivity disorder is characterized by a delay in cortical maturation. *Proceedings of the National Academy of Sciences of the United States of America, 104*(49), 19649-54.

Snyder, H. R., Miyake, A. & Hankin, B. L. (2015). Advancing understanding of executive function impairments and psychopathology: Bridging the gap between clinical and cognitive approaches. *Frontiers in Psychology, 6*, 328.

Snyder, H. R. & Munakata, Y. (2010). Becoming self-directed: Abstract representations support endogenous flexibility in children. *Cognition, 116*(2), 155-167.

Steinberg, L. (2005). Cognitive and affective development in adolescence. *Trends in Cognitive Science, 9*(2), 69-74.

Strauss, E., Sherman, E. M. & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. New York: Oxford University Press.

Thorell, L. B., Eninger, L. Brocki, K. C. & Bohlin, G. (2010). Childhood Executive Function Inventory (CHEXI): A promising measure for identifying young children with ADHD? *Journal of Clinical and Experimental Neuropsychology, 32*(1), 38-43.

Toplak, M. E., West, R. F. & Stanovich, K. E. (2013). Practitioner review: Do performance-based measures and ratings of executive function assess the same construct? *Journal of Child Psychology and Psychiatry, 54*(2), 131-143.

Torgesen, J. K., Wagner, R. K. & Rashotte, C. A. (2011). *TOWRE-2 Test of Word Reading Efficiency – Second Edition*. Austin, TX: Pro-Ed Publishing.

van der Sluis, S., de Jong, P. F. & van der Leij, A. (2007). Executive functioning in children, and its relations with reasoning, reading, and arithmetic. *Intelligence, 35*(5), 427-449.

Verbruggen, F. & Logan, G. D. (2008). Response inhibition in the stop-signal paradigm. *Trends in Cognitive Sciences, 12*(11), 418-424.

Wechsler, D. (2014). *WISC-V Technical and Interpretive manual*. Bloomington, MN: Pearson.

Wilhelm, O., Hildebrandt, A. & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology, 4*, 433.

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V. & Pennington, B. F. (2005). Validity of executive function theory of attention-deficit/hyperactivity disorder: a meta-analytic review. *Biological Psychiatry, 57*(11), 1336-1346.

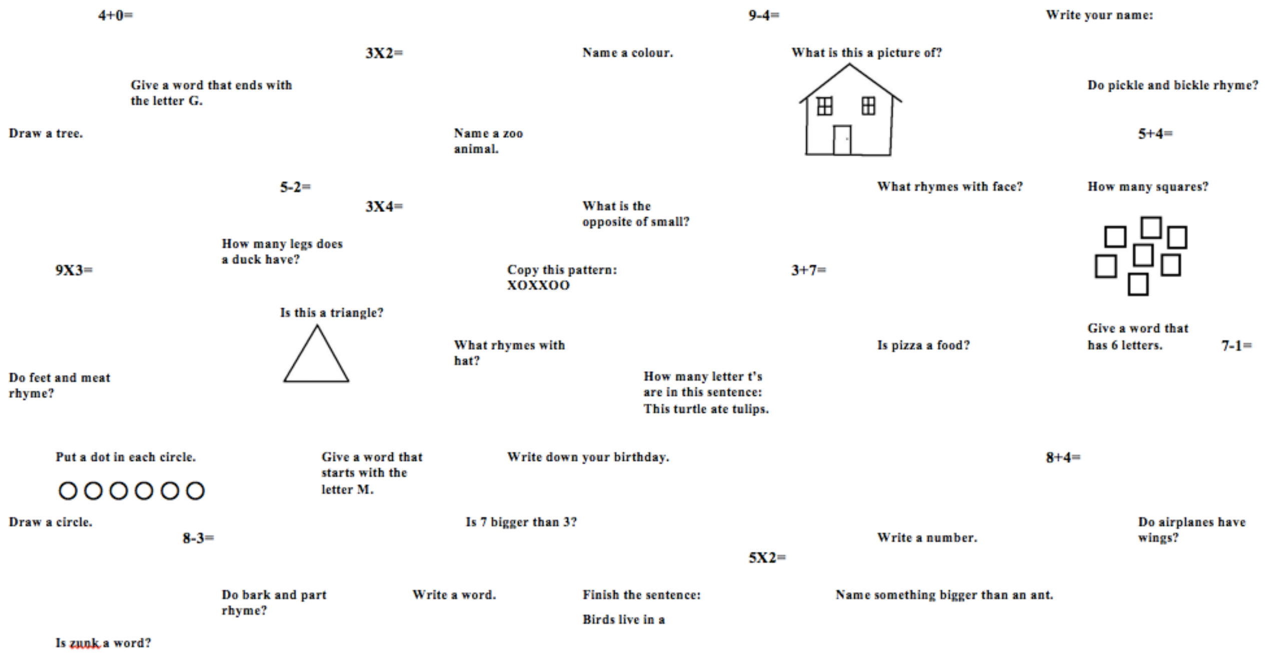
Xu, F., Han, Y., Sabbagh, M. A., Wang, T., Ren, X. & Li, C. (2013). Developmental differences in the structure of executive function in middle childhood and adolescence. *PLOS One, 8*(10), e77770.

Zelazo, P. D., Craik, F.I. & Booth, L. (2004). Executive function across the life span. *Acta*

Psychologica, 115(2-3), 167-183.

Zelazo, P. D. & Müller, U. (2002). Executive function in typical and atypical development. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 445-469). Malden, MA: Blackwell Publishers.

Appendix A: UPT and instructions



4+0=

9-4=

Write your name:

3X2=

Name a colour.

What is this a picture of?

Give a word that ends with the letter G.

Do pickle and bickle rhyme?

Draw a tree.

Name a zoo animal.

5+4=

5-2=

What rhymes with face?

How many squares?

3X4=

What is the opposite of small?

How many legs does a duck have?

9X3=

Copy this pattern: XOXXOO

3+7=

Is this a triangle?

Give a word that has 6 letters.

What rhymes with hat?

Is pizza a food?

7-1=

Do feet and meat rhyme?

How many letter t's are in this sentence: This turtle ate tulips.

Put a dot in each circle.

○○○○○○

Give a word that starts with the letter M.

Write down your birthday.

8+4=

Draw a circle.

Is 7 bigger than 3?

Do airplanes have wings?

8-3=

Write a number.

5X2=

Do bark and part rhyme?

Write a word.

Finish the sentence:

Birds live in a

Name something bigger than an ant.

Is zunk a word?

Instructions

"I would like you to complete the following worksheet. If you do not know the answer for any of the problems, just circle it and go on to the next problem. I cannot read any of the questions to you. Just do your very best, and when you are done, please bring the worksheet to me."

This is meant to be an unstructured task. The examiner should sit in another part of the room while the child is working on this task.

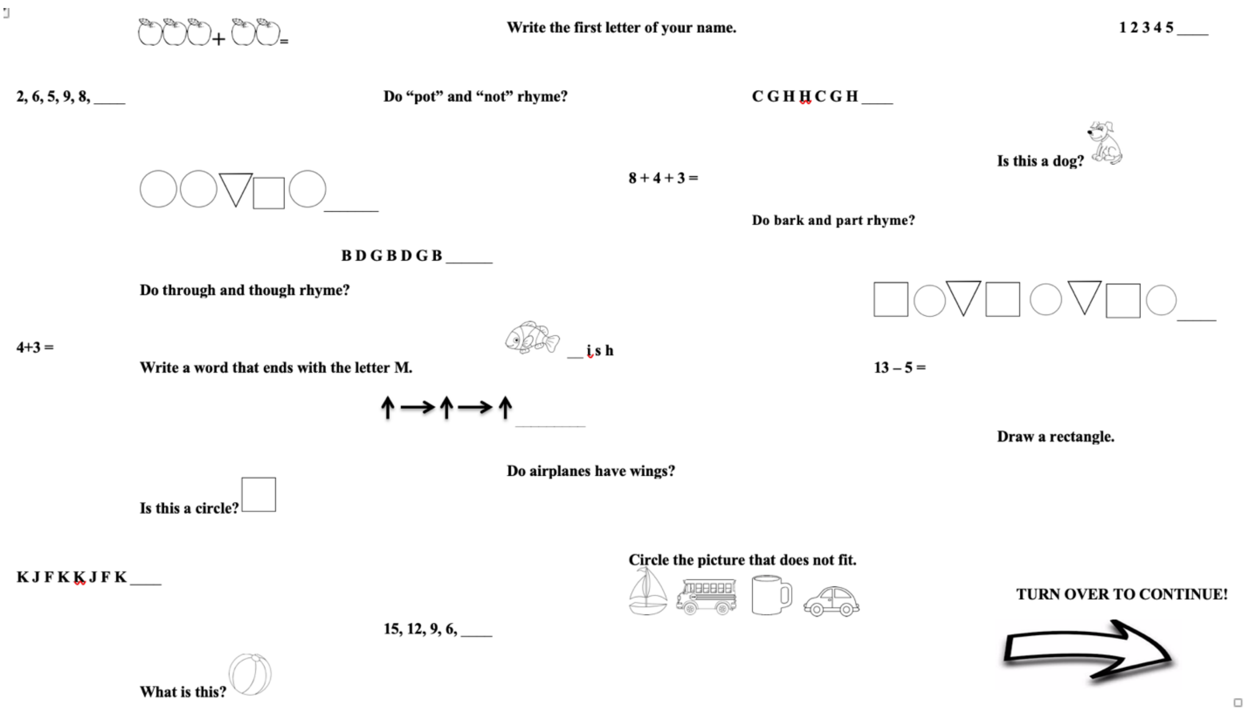
If the child displays oppositional behaviour (verbally or physically refusing to do the task e.g., I don't want to do this, or sitting with arms crossed and not picking up the pencil) the examiner may give ONE prompt:

"Please complete the worksheet. Remember you may circle items if you do not know the answer"

Note that the prompt was given and what occurred after the prompt (e.g., child continued to refuse to do task, child did task immediately after prompt, child refused to do task for a while but proceeded to task eventually or any thing else that happened) on the behavioural observations sheet.

Appendix B: UPT-2 and instructions

Side 1



+ =

Write the first letter of your name.

1 2 3 4 5 ____

2, 6, 5, 9, 8, ____

Do "pot" and "not" rhyme?

C G H H C G H ____

Is this a dog?

○○▽□○ _____

8 + 4 + 3 =

Do bark and part rhyme?

B D G B D G B ______

Do through and though rhyme?

□○▽□○▽□○ ___

4+3 =

Write a word that ends with the letter M.

__ i s h

13 – 5 =

↑→↑→↑ ______

Draw a rectangle.

Do airplanes have wings?

Is this a circle? □

Circle the picture that does not fit.

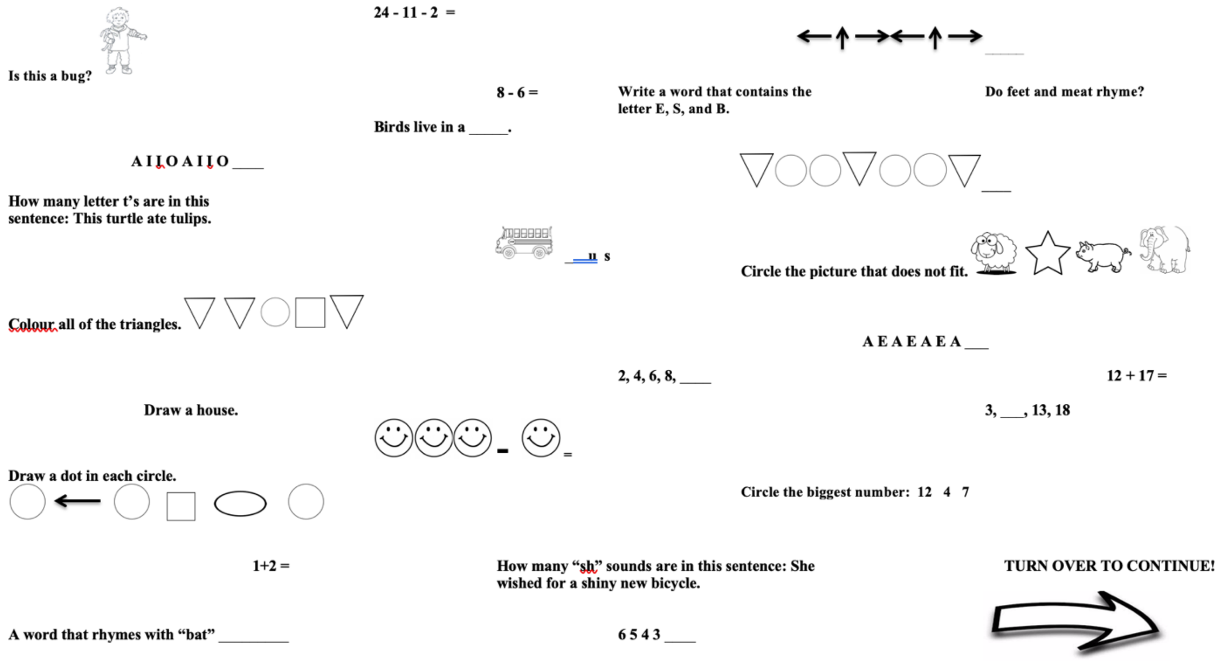
K J F K K J F K ____

TURN OVER TO CONTINUE!

15, 12, 9, 6, ____

What is this?

Side 2



24 - 11 - 2 =

←↑→←↑→ ____

Is this a bug?

8 - 6 =

Write a word that contains the letter E, S, and B.

Do feet and meat rhyme?

Birds live in a _____.

A I I O A I I O ____

▽○○▽○○▽ __

How many letter t's are in this sentence: This turtle ate tulips.

__u s

Circle the picture that does not fit.

Colour all of the triangles. ▽▽○□▽

A E A E A E A ___

2, 4, 6, 8, ____

12 + 17 =

Draw a house.

3, ___, 13, 18

☺☺☺ - ☺ =

Draw a dot in each circle.

○ ← ○ □ ⬭ ○

Circle the biggest number: 12 4 7

1+2 =

How many "sh" sounds are in this sentence: She wished for a shiny new bicycle.

TURN OVER TO CONTINUE!

A word that rhymes with "bat" _________

6 5 4 3 ____

Instructions

"I would like you to complete the following worksheet. If you do not know the answer for any of the problems, just circle it and go on to the next problem. I cannot read any of the questions to you. Just do your very best, and when you are done, please bring the worksheet to me."

This is meant to be an unstructured task. The examiner should sit in another part of the room while the child is working on this task.

If the child displays oppositional behaviour (verbally or physically refusing to do the task e.g., I don't want to do this, or sitting with arms crossed and not picking up the pencil) the examiner may give ONE prompt:

"Please complete the worksheet. Remember you may circle items if you do not know the answer"

Note that the prompt was given and what occurred after the prompt (e.g., child continued to refuse to do task, child did task immediately after prompt, child refused to do task for a while but proceeded to task eventually or any thing else that happened) on the behavioural observations sheet.