

METACOGNITIVE RATINGS OF TASK DIFFICULTY, EFFORT EXERTED, EFFORT  
REQUIRED AND AFFECTIVE EXPERIENCE OF EFFORT ON AN UNSTRUCTURED  
PERFORMANCE TASK IN A COMMUNITY SAMPLE OF CHILDREN

KAITLYN MARIE BUTTERFIELD

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

JULY 2022

## Abstract

Metacognitive ratings of effort are typically assessed by asking participants to indicate their mental effort on a performance-based task. Executive functions enable problem solving and goal attainment. Historically, EFs have been assessed using performance-based measures and rating scales. Research has shown a lack of association between these two methods. One framework used to understand this difference is the structure provided on performance-based measures and not provided on rating scales. This study investigated the role of structure by examining a novel Unstructured Performance-based Task (UPT-2) and metacognitive ratings of effort. Ninety-eight children between the ages of five and 11 years ( $M = 9.33$ ,  $SD = 1.75$ , 47 females) from an independent school were recruited in Fall 2018. Significant associations emerged between the UPT-2, EF tasks and ratings, academic abilities, and metacognitive ratings of effort. The rating of effort required emerged as a predictor of performance on the UPT-2. Results suggest the UPT-2 may be a promising measure to assess EF-related difficulties and provide an understanding of children's behaviors in unstructured environments. Further, these findings consider the importance of specific reference points in rating scales. Finally, developmental sensitivity must be considered in future UPT-2 research to better understand the contribution of metacognitive ratings of effort and performance on an unstructured task.

## Acknowledgements

I would like to thank the families, teachers, and students who participated in this study. Without you, this project would not have been possible. You have taught us so much.

I am grateful to Maggie Toplak for bringing me into the program of my dreams. Your support, guidance, and endless (and unbelievably helpful) analogies were the driving force behind this project. Elizabeth and Rachael, thank you for all of your support and willingness to chat at just about any hour of the day. This project would not exist without the foundation built by Justine and Elizabeth's prior work on the UPT – thank you for your hard work and dedication to this important research.

This project would not have been possible without the generosity of Dr. Rhonda Martinussen. Thank you for your willingness to collaborate on this study and for your incredibly helpful feedback, kindness, and direction throughout.

To my committee members, Drs. Mary Desrocher and John Eastwood, I have so much gratitude for your willingness to support my research.

I am privileged to be a part of the most extraordinary cohort – Sara, Jenna, Kate, Nisha, Meaghan, Teresa, Paolina, Ethan, Megis, and Katherine – you are forces to be reckoned with.

Finally, thank you to my family for your unwavering support. “Butterfield’s never give up”.

## Table of Contents

Abstract.....	ii
Acknowledgements.....	iii
Table of Contents.....	iv
List of Tables.....	vi
List of Figures.....	vii
List of Acronyms.....	viii
Introduction.....	1
Performance-Based Measures and Behavioral Rating Scales of Executive Function.....	1
Degree of Structure.....	2
Metacognitive Ratings of Task Difficulty, Effort Exerted, Effort Required, and Affective Experiences of Effort.....	5
Overview of the Current Study.....	7
Hypotheses.....	8
Method.....	10
Participants.....	10
Measures.....	11
Procedure.....	15
Statistical Plan.....	15
Results.....	16
Descriptive Statistics and Scoring on the UPT-2.....	16
UPT-2 Total Correct Items.....	17
UPT-2 Total Complete Items.....	19
Time to Complete.....	20
Descriptive Statistics and Scoring on the Metacognitive Ratings of Effort.....	22
Task Difficulty.....	22
Effort Exerted.....	23
Effort Required.....	23
Affective Experience of Effort.....	24
Correlations between Age, UPT-2 Performance, and UPT-2 Metacognitive Ratings.....	25
Correlates of the Unstructured Performance Task – 2 <sup>nd</sup> Version.....	27
Performance-based Tasks and a Rating Scale of Executive Function.....	27
Academic Abilities.....	29
Metacognitive Ratings of Effort.....	31
Correlates within Metacognitive Ratings of Effort.....	32
Predictors of the Unstructured Performance Task – 2 <sup>nd</sup> Version.....	34
Metacognitive Ratings.....	34
Performance-based Tasks of Executive Function and Effort Required.....	35
Behavioral Rating Scale of Executive Function and Effort Required.....	36
Predictors of Reading Abilities.....	37
Predictors of Math Abilities.....	38
Discussion.....	40
Associations among the UPT-2, Performance-based tasks of Executive Function, and Behavioral Rating Scale.....	41
Associations among UPT-2 Performance and Academic abilities.....	42
Associations among UPT-2 Performance and Age.....	45

Metacognitive Ratings of Effort.....	43
Associations among UPT-2 Performance and Metacognitive Ratings.....	44
Multiple Regression predicting UPT-2 performance from Metacognitive Ratings....	45
Hierarchical Regression predicting UPT-2 performance.....	46
Hierarchical Regression predicting Academic Abilities.....	47
Implications of Metacognitive Ratings of Effort.....	48
Implications for Clinical Assessment and Education.....	49
Limitations.....	50
Conclusion.....	52
References.....	53
List of Appendices.....	64
Appendix A.....	65
Appendix B.....	66

## List of Tables

Table 1: Frequency Distribution of Gender (Male, Female) and Grade (1-6) (N=98).....	11
Table 2: Potential and actual range, means, standard deviations, skewness, and kurtosis indices for all variables.....	16
Table 3: Spearman correlations between all variables and age.....	26
Table 4: Spearman Correlations between UPT-2 variables, Performance-based EF Tasks, and a Rating Scale of EF.....	28
Table 5: Spearman Correlations between UPT-2 variables and Academic Abilities.....	30
Table 6: Spearman Correlations between UPT-2 variables and Metacognitive Ratings of Effort...32	32
Table 7: Spearman correlations within metacognitive ratings of effort (task difficulty, effort exerted, effort required, and affective experience of effort).....	33
Table 8: Spearman correlations within metacognitive ratings of effort (task difficulty, effort exerted, effort required, and affective experience of effort) and UPT-2 composite performance in low vs. high scorers.....	34
Table 9: Regression results predicting UPT-2 Composite Performance from metacognitive ratings of effort.....	35
Table 10: Hierarchical Regression Analysis for UPT-2 Composite Performance (N=98).....	36
Table 11: Regression results predicting UPT-2 Composite Performance from an EF rating scale and effort required rating.....	37
Table 12: Hierarchical Regression Analysis for Reading Abilities (N=98).....	38
Table 13: Hierarchical Regression Analysis for Math Abilities (N=98).....	39
Table 14: Summary of Results.....	39

## List of Figures

Figure 1a. Distribution of correct items on the UPT-2 across the entire sample.....	18
Figure 1b. Distribution of correct items on the UPT-2 by grade.....	18
Figure 2a. Distribution of complete items on the UPT-2 across the entire sample.....	19
Figure 2b. Distribution of complete items on the UPT-2 by grade.....	20
Figure 3a. Distribution of Time to Complete the UPT-2 across the entire sample.....	21
Figure 3b. Distribution of Time to Complete the UPT-2 (in seconds) across the entire sample...	21
Figure 4. Distribution of Task Difficulty across the entire sample.....	22
Figure 5. Distribution of Effort Exerted across entire sample.....	23
Figure 6. Distribution of Effort Required across the entire sample.....	24
Figure 7. Distribution of Affective Experience of Effort across entire sample.....	25
Figure 8. Scatter plots of correlations between composite performance on the UPT-2 and performance-based tasks and a rating scale of EF.....	29
Figure 9: Scatter plots of correlations between composite performance on the UPT-2 and measures of academic achievement (WJTA and TOWRE) .....	31

## List of Acronyms

**BDEFS-CA**= Barkley Deficits in Executive Function Scale – Children and Adolescents Short Form

**BRIEF**= Behavioral Rating Inventory of Executive Function

**BSP**= Barriers & Strategies Protocol

**EF**= Executive Function

**TMT**= Trail-Making Test

**TOWRE2**= Tests of Word Reading Efficiency -2<sup>nd</sup> Edition

**UPT**= Unstructured Performance Task-1<sup>st</sup> Version

**UPT-2**= Unstructured Performance Task-2<sup>nd</sup> Version

**WJ-IV**= Woodcock-Johnson Tests of Intelligence -4<sup>th</sup> Edition





## **Metacognitive Ratings of Task Difficulty, Effort Required and Affective Experience on an Unstructured Performance Task (UPT-2) in a Community Sample of Children**

It has been consistently reported that performance-based measures of executive function and rating scale measures of executive function show a low to modest correlation (Bodnar et al., 2007; Mahone et al., 2002; McAuley et al., 2010; Toplak et al., 2013). Executive functions (EFs) generally refer to higher-order abilities used for problem-solving and goal-directed behavior (Anderson, 2002; Diamond, 2013; 2020; Miyake et al., 2000; Pennington & Ozonoff, 1996) and executive function skills have been measured using performance-based tasks (Diamond, 2013; 2020; Zelazo et al., 2013) and rating scales (Gioia et al., 2008; 2015). This finding of the low to modest correlations between tasks have been of considerable interest, both conceptually and empirically (Toplak et al., 2013), but also from the perspective of understanding the conditions under which these measures will display higher and lower correlations.

The issue of task structure has been described as an important consideration for the effect size of these correlations. The Unstructured Performance Task (UPT) was designed to minimize task structure and examiner direction (Ledochowski et al., 2019; Wanstall, 2019). An additional way to understand how performance and ratings may be related is by using metacognitive ratings to assess how participants monitor and control their cognitive resources during a task.

Specifically, metacognitive ratings of task difficulty, effort required, effort exerted, and affective experience of effort were examined in the current study as correlates of performance on the UPT in a community sample of children.

### **Performance-Based Measures and Behavioral Rating Scales of Executive Function**

The assessment of performance-based measures of executive function reflects critical skills and abilities that support academic development and are critical to supporting cognitive

and emotional development (Diamond & Ling, 2016). However, while these measures seem to assess important competencies related to efficiency and information capacity, Gioia et al. (2008) posits that performance-based measures may fail to capture other important aspects of EF:

Individuals with substantial executive dysfunction can often perform adequately on well-structured tests when the examiner is allowed to cue and probe for more information, relieving the individual of the need to be appropriately inhibited, flexible, strategic in planning, and goal directed. (p. 180).

That is, performance-based measures of executive function consist of tasks conducted under highly standardized conditions, unlike many of the less structured situations that children would typically encounter in their daily lives. Performance on tasks such as the Stroop Test (Jensen & Rohwer, 1966; MacLeod, 1991; Stroop, 1935), the Wisconsin Card Sorting Test (WCST; Heaton et al., 1993), and tests of verbal fluency (Strauss et al., 2006) quantify the examinee's accuracy and response time. Given these considerations with respect to performance-based measures, rating scales were developed to be an ecologically valid indicator of competence in complex, every day, problem-solving situations (Gioia et al., 2000; O'Brien et al., 2021; Toplak et al., 2013). For example, the Behavioral Rating Inventory of Executive Function (BRIEF; Gioia et al., 2000) is a rating scale that involves informants reporting on how well individuals can manage and organize themselves on real-world tasks. However, the small to modest associations between performance-based and rating measures of EF have raised important conceptual and empirical questions, including understanding conditions under which these associations may be increased or decreased.

### **Degree of Task Structure and Direction**

Task structure and direction from the examiner have been suggested as important characteristics of performance-based measures, including executive function tasks (Stanovich, 2009). Stanovich (2009) postulates that:

Executive processes are misnamed in the psychological literature. Executive functioning measures are nothing of the kind—at least as most people would understand the word “executive”. These tasks might instead be better termed measures of *supervisory* processes. They assess the ability to carry out the rules instantiated not by internal regulation (*true* executive control) but by an external authority that explicitly sets the rules and tells the subject what constitutes maximal performance. The subject does not set the agenda in these tasks (as is the case in many tasks in the rational thinking and critical thinking literatures) but instead attempts to optimize criteria explicitly given to them. The processes assessed by such tasks do involve algorithmic-level decoupling (which is why they are so highly related to fluid intelligence), but they are supervisory in nature—decoupling is used to screen out distracting stimuli and make sure the externally-provided rule remains the goal state. (p. 66)

Conceptually, performance-based measures rely on maximal-optimal performance, such that the task is interpreted with direction and instruction from an examiner. These instructions enable the examinee to maximize their performance because they are explicitly told how to do so. Performance-based measures of executive function tend to have explicit administrative instructions for examiners including: set-up of physical environment to minimize distraction, specific instruction for the examiner on how to complete a task and what types of responses are required from the examinee (sample problems with feedback may be completed), corrective feedback during administration if the task is misunderstood, pacing in control by the examiner, and possibility for redirecting an examinee after long delays or a specific time limit has elapsed.

In contrast, rating measures operationalize “typical performance” by asking questions about cognitive and executive skills that are experienced during everyday situations. These typical performance situations are unconstrained, such that no overt instruction is provided to the examinee to allow for maximal performance. Rating scales of executive function are unlikely to address the physical environment where the examinee completes the ratings. While there typically are instructions for the examinee to complete the questions, it is not necessary that they are presented by the examiner. Thus, the likelihood of rating scales being self-administered means that there is no corrective feedback provided upon completion of any question. Unlike

performance-based tasks, rating scales are self-paced and there are no specific time limits for completion of questions; it is even possible for an examinee to complete questions on multiple occasions. The reference point for these rating scales may be very general, as they may not refer to any specific situation or point in time. Alternatively, metacognitive ratings of effort are typically conducted with reference to a specific task. Given that metacognitive ratings have a more specific reference point, these ratings would be expected to correlate more strongly with performance on the UPT-2 than the behavioural rating scales.

Having in mind the distinction between maximal and typical performance distinction made by Stanovich (2009), we developed the Unstructured Performance Task (UPT) in our lab. Its conceptualization involved operationalizing the maximal-typical distinction, such that successful task completion relies more heavily on pacing and regulation by the examinee than on explicitly set rules and guidance by an examiner. The UPT was designed to be a relatively easy task where the participant is asked to complete several items independently and inform the examiner when the task has been completed (Ledochowski et al., 2019). Despite one of its design features being that the UPT was intended to be a relatively easy task, its difficulty stems from being tedious and requiring time, effort, and motivation to complete with little direction from an examiner. Following Wanstall's (2019) first study on the UPT, the UPT-2 was developed as a performance-based measure to reduce the task structure and direction from the examiner to determine whether these characteristics may at least partly explain the low to modest correlations between performance-based tasks and behavioral rating scales of EF.

Initial work on the UPT has demonstrated that performance on this task was uniquely predicted by both performance-based tasks of EF and a behavioral rating scale of EF in a sample of children with and without ADHD, suggesting that the UPT captures EF-related behaviour and

performance under less structured conditions (Ledochowski et al., 2019). Following a revision of the UPT-2, Wanstall et al. (in preparation) found that performance on the UPT-2 was predicted by performance-based EF tasks in a community sample but did not replicate the unique prediction with a behavioral rating scale. Additionally, they reported a significant association between EF performance-based tasks, a behavioral rating scale, and performance on the UPT-2. Put together, correlations of the UPT-2 with both performance-based measures and a behavioral rating scale of executive function indicate that performance demands (maximal) and less structure (typical) are characteristic of this novel task.

Given the unique design feature of having participants complete relatively easy test items with little direction from an examiner, metacognitive ratings of the UPT-2 were of considerable empirical interest. Thus, in addition to examining performance on this task, participants were asked to provide metacognitive ratings of effort to examine whether perceived task difficulty, effort exerted, effort required, and affective experience of effort would be related to their performance. Given the design features of the UPT-2 as a relatively easy task that requires participants to independently complete this task, the novel contribution of the current study was to examine metacognitive correlates of UPT-2 performance.

### **Metacognitive Ratings of Task Difficulty, Effort Exerted, Effort Required and Affective Experience of Effort**

The metacognitive literature points to two processes that allow an individual to monitor their performance and control the allocation of cognitive resources (Ackerman & Thompson 2017). There is an integral relationship between (1) object-level and (2) online monitoring and allocation of resources, such that object-level processes are comprised of “perceiving,

remembering, classifying, and deciding”, and meta-level processes include the monitoring of object-level processes (Ackerman & Thompson, 2017).

Metacognitive ratings of performance can provide a complementary set of analyses in order to further understand the correlations between EF tasks and ratings. For example, the use of metacognitive ratings of effort provides indicators that are expected to be related to actual performance (Bjork et al., 2013). With respect to the assessment of effort exerted during a cognitively demanding task, at least three different aspects of effort have been defined in the literature; 1) *how difficult the task was* (task difficulty), 2) *how hard I tried* (effort exerted), and 3) *how taxed I felt* (effort required; Hsu et al., 2017). Research has found an association between metacognitive ratings of effort and performance; a 2017 study reported an increase in subjective ratings of task difficulty with a significant decrease in learning success (Korbach et al., 2017).

In a recent study, ratings of anticipated, real-time and recalled subjective effort were compared on a sustained attention task (less cognitively demanding) and a working memory task (more cognitively demanding) in a sample of undergraduate students (Bambrach et al., 2019). Participants who completed the working memory task anticipated significantly more cognitive effort than they reported actually experiencing after the task. However, participants who completed the sustained attention task anticipated significantly less cognitive effort than they reported actually experiencing after the task. Given that the sustained attention task seemed to be a less difficult task, it was somewhat surprising that despite its low difficulty, participants reported experiencing more cognitive effort to complete this task. Relating this to the UPT-2, which was designed to be a less cognitively demanding task, our predictions were more similar to sustained attention findings in the Bambrach et al. (2019) study.

One would expect that the metacognitive ratings would track performance patterns on the UPT-2. For example, an individual with high accuracy and completion on the UPT-2 may rate the task as relatively easy (given that it was designed to be an easy task) and requiring some effort (considering its tediousness). Alternatively, in an individual with low accuracy and completion on the UPT-2, their metacognitive ratings of effort can point to potential contributors to their performance. More specifically, the metacognitive ratings of effort can help to qualitatively differentiate between an individual with a low performance score on the UPT-2 who rates the task as relatively easy and trying their best from an individual with low performance who rated the task as relatively easy and not trying their best. It has been reported that the real-time ratings of effort required is negatively correlated with performance accuracy on a working memory task in students at-risk for ADHD, but not in the non-at-risk group (Hsu et al., 2017). These findings suggest that metacognitive ratings of effort required may correlate differently with objective task performance depending on the sample, at least on a working memory task.

### **Overview of the Current Study**

In the current study, participants completed the UPT-2 and answered the following metacognitive ratings following the task: (1) task difficulty (*how hard was the task?*), (2) effort exerted (*how hard did you try?*) (3) effort required (*how much brainpower did you use?*), and (4) affective experience of effort (*how did using brainpower make you feel?*). In addition, performance-based EF tasks, a rating scale of EF, and academic achievement measures were examined. The following hypotheses were made in relation to previous work on the UPT and UPT-2, however, findings from the current study may differ considering the clinical sample in Ledochowski et al. (2019) and the use of a previous UPT version in part of Wanstall (2019).



## *Hypotheses*

1. Executive Function Ratings and Performance-Based Tasks: Performance on the UPT, based on accuracy and completion, has been shown to be significantly correlated with performance-based measures of executive function (EF tasks; Stroop Test and Trail Making Test) and ratings of executive function (EF ratings; BDEFS-CA parent ratings) in a clinical sample of children with and without ADHD (Ledochowski et al., 2019). In contrast, the UPT, based on accuracy and completion, was only correlated with EF tasks (and not EF ratings) in a community sample of children (Wanstall, 2019). It was hypothesized that the findings from Wanstall (2019) would be replicated in the community sample of children in this study. In addition, it was also expected that performance-based tasks and the behavioral rating scale of EF would not be significantly correlated (see Toplak et al., 2013 for a review).
2. Academic Abilities: Performance on the UPT-2 was expected to be significantly correlated with academic abilities (reading and math achievement) in a community sample of children. Children with higher academic achievement were expected to have higher UPT-2 performance than children with lower academic abilities. This would replicate the findings of Wanstall et al. (in preparation).
3. Correlations within Metacognitive Ratings of Effort:
  - a. The metacognitive ratings of effort can be categorized as task-based (task difficulty, effort required) and examinee-based (effort exerted, affective experience of effort). It was therefore predicted that task-based ratings would be significantly positively correlated (task difficulty and effort required), and examinee-based ratings would be significantly positively correlated (effort exerted and affective experience of effort).

- b. The amount of effort exerted should depend on task difficulty, thus it was hypothesized that task difficulty and effort exerted would be significantly positively correlated.
- c. Finally, we predicted that children would exert the amount of effort that was required, therefore these ratings were also expected to be significantly positively correlated.

#### 4. Metacognitive Ratings and UPT-2 Performance

- a. As the UPT-2 was designed to be a relatively easy task, it was hypothesized that performance on the UPT-2 would be negatively correlated with the task difficulty rating (Maynard & Hakel, 1997; Li et al., 2007). Those participants who consider the UPT-2 to be a difficult task (higher rating), possibly due to its tediousness, were expected to do less well.
- b. Despite the UPT-2 being designed as a relatively easy task, it was expected that performance on the UPT-2 would be correlated with the effort exerted rating, as a participants who exerts little to no effort would not be expected to perform as well as a participant who exerted maximal effort.
- c. It was predicted that performance on the UPT-2 would be significantly correlated with the effort required rating. That is, better performance on the UPT-2 would be correlated with less effort required.
- d. Given the community sample of high-achieving students in the current study, it was hypothesized that there would be no significant relation between UPT-2 performance and the affective experience of effort.

#### 5. Regression Analyses

- a. UPT-2 and Metacognitive Ratings of Effort: It was hypothesized that when entered alone, metacognitive ratings of effort would enter as a significant predictor of performance on the UPT-2.
- b. UPT-2 and Performance-Based Tasks: It was hypothesized that the performance-based tasks of EF and effort required rating would significantly predict UPT-2 performance, after statistically controlling for age.
- c. UPT-2 and the Rating Scale of EF: It was hypothesized that parent-rated EF and the effort required rating would significantly predict UPT-2 performance.
- d. UPT-2 and Reading Abilities: Wanstall (2019) found that age and the UPT-2 performance predict reading abilities, whereas performance-based EF tasks and the rating scale of EF did not. We hypothesized that only age, UPT-2 performance, and the effort required rating would predict reading abilities, but not performance-based EF tasks or rating scales.
- e. UPT-2 and Math Abilities: Wanstall (2019) found that age, UPT-2 composite performance, and the EF rating scale predict math abilities, whereas performance-based EF tasks do not. We hypothesized that after entering age, UPT-2 composite performance, and the effort required rating, the rating scale of EF would emerge as a significant predictor.

## **Method**

### **Participants**

Ninety-eight children between the ages of five and 11 years ( $M = 9.33$ ,  $SD = 1.75$ , 47 females) were recruited in Fall 2018. This study was embedded within a larger study that was conducted at an independent school, housed within a major university in a metropolitan city. All

students were in grades 1 through 6. The frequency of students in each grade and gender are presented in Table 1. Age and gender were relatively well balanced in these groups.

**Table 1**

*Frequency Distribution of Gender (Male, Female) and Grade (1-6) (N=98)*

	<b>Frequency</b>	<b>Percent</b>
Gender		
Male	51	52
Female	47	48
Grade		
1	13	13.3
2	17	17.3
3	18	18.4
4	16	16.3
5	14	14.3
6	20	20.4

## Measures

### *The Unstructured Performance Task – Version 2 (UPT-2)*

The UPT assesses performance while minimizing structured imposed by the task and the examiner (Ledochowski et al., 2019). The UPT-2 was an adaptation to minimize skew of performance and ceiling effects (Wanstall, 2019). The UPT-2 was administered to children individually in a quiet room by trained facilitators. The UPT-2 contained 50 simple questions (25 on each page) in the domains of math, reading, general knowledge, and rote copying. The task was presented on a double-sided large sheet of paper (11x17 inches). Questions were presented randomly on the page, were not numbered, and did not contain any prompts. Participants completed the task at their own pace. Facilitators were trained to provide little to no instruction or aid on the task. Instructions were read to participants as follows: “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 move on to the next problem. I cannot read any of the questions to you. Do your

very best, and when you are done, please bring the worksheet to me”. After the initial instructions, the facilitator remained silent and physically distanced in the room, and any questions from the participant were answered with “do what you think is best”. Time limits were not given to participants; however, they were timed and generally did not take more than 15 minutes to complete the task. The UPT-2 was scored across four domains: accuracy (number of correct items), completeness (number of items attempted, regardless of accuracy), missed items (items not attempted and not circled as unknown), and unknown items (items circled as unknown). The current study used Wanstall et al.’s (in preparation) scoring algorithm for the UPT-2 that would simultaneously account for accuracy and completion of the items, weighting completion more heavily than accuracy to better capture the self-directive aspect of the UPT-2. This UPT-2 combined score was calculated as follows:  $\text{UPT-2 correct} + 2 * (\text{UPT-2 complete})$ . Appendix A displays a visual image of the UPT-2.

### ***Metacognitive Ratings of Effort***

This measure was administered immediately after participants completed the UPT-2. The metacognitive ratings of effort were presented on a single sided sheet of paper. Four child-friendly questions were used to assess task difficulty, effort exerted, effort required and affective experience of effort. Each question has its own scale; (1) “How hard was this activity for you?” is scored on a 5-point Likert scale where participants were asked to circle one of five illustrations of a stick person carrying a backpack (e.g., 1-carrying a light backpack, 5-carrying a heavy backpack); (2) “How hard did you try on this activity?” is scored on a 5-point Likert scale (1-I did not try at all, 5-I tried my best); (3) “How much brainpower did you use on this activity?” is scored on a 3-point Likert scale where participants were asked to mark one of three illustrations of a brain (1-brain is sleeping, 3-brain is lifting heavy weights); (4) “How did using brainpower

make you feel?” is scored on a 5-point Likert scale where participants were asked to circle one of five illustrations (1-a red face with its tongue out, 5-a yellow face smiling with teeth). See Appendix C for a visual image of the metacognitive ratings of effort.

### *Academic performance*

**Woodcock Johnson Tests of Achievement – Fourth Edition (WJ-IV).** The WJ-IV is a commonly used standardized measure of academic achievement in children and adolescents. Participants were administered two subtests from the WJ-IV: Math Fluency and Math Calculation. The Math Fluency subtest is a timed task where participants are asked answer as many simple math problems (i.e., addition, subtraction, multiplication) as possible in three minutes. The Math Calculation includes items that gradually increase in difficulty. This subtest is not timed, and participants are asked to solve as many math problems as they can. According to Villarreal (2015), internal consistency is excellent with a Cronbach’s alpha range from .84 to .94. A combined z-score (not age-corrected) was generated from raw scores based on accuracy of performance, with a higher score indicating better performance. Both subtests were administered in a group format.

**Test of Word Reading Efficiency – Second Edition (TOWRE-2).** The TOWRE-2 was developed by Torgesen and colleagues (2012) as a standardized measure of sight word recognition and phonemic decoding in children and adults. Participants were administered two subtests from the TOWRE-2: the Sight Word Reading Efficiency and the Phonemic Decoding subtests. Both subtests were administered to individual participants in a quiet room. The Sight Word Reading Efficiency subtests assesses the number of real printed words that participants accurately read within 45 seconds. The Phonemic Decoding subtest measures the number of pronounceable and printed non-words that participants accurately decoded within 45 seconds.

Doty et al. (2015) has shown high internal consistency for the TOWRE-2, with Cronbach's alpha ranging from .90 to .98. To create an overall reading score, a combined z-score was generated from raw scores, with higher scores indicating better performance.

### *Executive Function Rating Scale*

**Barkley Deficits in Executive Functioning Scale – Children and Adolescents Short Form (BDEFS-CA; Barkley, 2012).** The BDEFS- CA short form is a 20-item rating scale completed by parents and used to assess child and adolescent executive functioning. The questions assess time management, organization and problem solving, self-restraint, self-motivation, and self-regulation of emotions. This measure has been found to be both reliable and valid (Barkley, 2012), with internal consistency of .95 for the 20-item form. An overall total score was derived based on summing the items, and higher scores indicated more difficulties in executive functioning.

### *Executive Function Performance-Based Tasks*

**Trail Making Test (TMT; Reitan, 1955; 1958).** 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 paper and pencil task is administered by an examiner. In Part A of this task, participants were 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 dependent measure for this task was calculated by subtracting completion time on Part A (processing speed) from completion time on Part B (set-shifting).

**Stroop Color-Word Test (Golden, 1978).** The Stroop Test 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 dependent 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).

### **Procedure**

All children required written consent of a primary caregiver prior to providing their own verbal assent. Participants completed tasks and measures under the direction of an examiner.

*Statistical plan.* Correlations between scores and indices within the UPT-2 across the entire sample were examined. Additionally, correlations among age, performance-based measures of EF, a rating scale of EF, academic abilities, metacognitive ratings of effort, and the UPT-2 were examined across the entire sample. A median split of UPT-2 performance was used to further explore the association between metacognitive ratings of effort and performance on the UPT-2. A multiple regression was conducted to determine whether metacognitive ratings of effort predicted performance on the UPT-2. Hierarchical regressions were conducted to determine whether performance-based tasks and a rating scale of EF predicted performance on



the UPT-2 after controlling for age. Finally, a set of hierarchical regressions were conducted to further Wanstall's (2019) findings on academic achievement and the UPT-2.

## Results

### Descriptive Statistics and Scoring on the UPT-2

Descriptive statistics of the means and ranges of all study measures are shown in Table 2. There was clear evidence of a negative skew on the UPT-2, when looking at both correct and complete scores. This is consistent with findings from Ledochowski et al. (2019) and Wanstall (2019). As such, non-parametric analyses were used in this study, namely Kruskal-Wallis tests as well as Spearman correlations.

**Table 2**

*Potential and actual range, means, standard deviations, skewness, and kurtosis indices for all variables*

Variable	Potential range	Actual range	Mean( <i>SD</i> )	Skewness	Kurtosis
<b><u>UPT-2 Performance and Ratings</u></b>					
UPT-2 Total Correct	0 to 50	4 to 50	37.87(11.44)	-1.26	0.79
UPT-2 Total complete	0 to 50	20 to 50	45.38(6.19)	-2.03	4.05
UPT-2 Composite Performance	0 to 150	38 to 150	121.11(27.98)	-1.34	1.10
UPT Task Difficulty Rating (5=most difficult)	1 to 5	1 to 5	2.08(0.99)	0.63	-0.29
UPT Effort Exerted (5=tried my best)	1 to 5	1 to 5	4.13(1.10)	-1.15	0.51
UPT Effort Required (3=most brainpower)	1 to 3	1 to 3	1.97(0.62)	0.19	-0.33
UPT Affective Rating of Effort Required (5=very good)	1 to 5	1 to 5	3.73(0.98)	-0.31	-0.05
<b><u>Executive Function Ratings and Performance-Based Tasks</u></b>					
BDEFS-CA (parent)	20 to 80	20 to 55	33.63(8.86)	0.58	-0.52
Stroop Interference Time	N/A	-2 to 119	39.02(21.10)	1.45	3.09

TMT Part B minus Part A Time	N/A	-7 to 296	69.12(45.99)	1.79	5.58
<b><u>Academic Tasks</u></b>					
TOWRE-2 Phonemic Decoding Efficiency Raw Score	0 to 66	2 to 60	35.83(13.66)	-0.31	0.82
TOWRE-2 Sight Word Efficiency Raw Score	0 to 108	17 to 107	66.19(19.70)	-0.57	-0.24
TOWRE-2 Combined z-Score	N/A	-2.49 to 1.92	0(.97)	-0.47	0.48
WJTA-IV Math Calculation Raw Score	N/A	10 to 46	36.74(98.55)	9.80	96.59
WJTA-IV Math Fluency Raw Score	0 to 180	7 to 160	78.99(136.43)	6.44	42.27
WJTA-IV Combined z-Score	N/A	-1.94 to 2.67	.01(.96)	0.27	-0.26

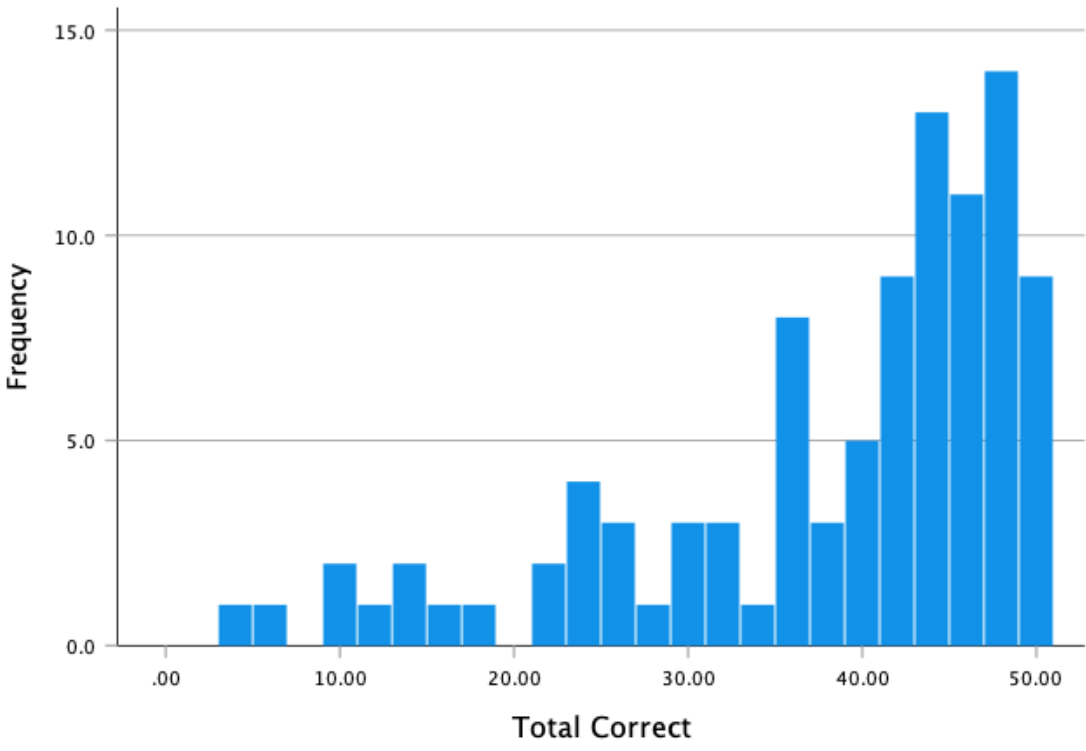
*Note.* \*  $p < .05$ ; \*\*  $p < .01$ . UPT=Unstructured Performance Task, WJTA=Woodcock-Johnson Tests of Achievement, TOWRE=Test of Word Reading Efficiency, BDEFS-CA=Barkley Deficits in Executive Functioning Scale.

### ***UPT-2 Total Correct Items***

Across the entire sample, participants answered an average of 37.87 ( $SD=11.44$ ) items out of 50 correctly on the UPT-2. The number of correct items increased developmentally, with a significant effect of grade on total correct items on the UPT-2 [ $\chi^2(170)=234.20, p < .001$ ]. There were no significant differences between males ( $M=37.22, SD=11.42$ ) and females ( $M=38.57, SD=11.54$ ) on UPT-2 total correct scores [ $W=2438.5, p=.43$ ]. Cronbach's alpha for the total items correct on the UPT-2 revealed good internal consistency ( $\alpha=.96$ ). See Figure 1a and 1b for distributions of correct items across the entire sample and by grade.

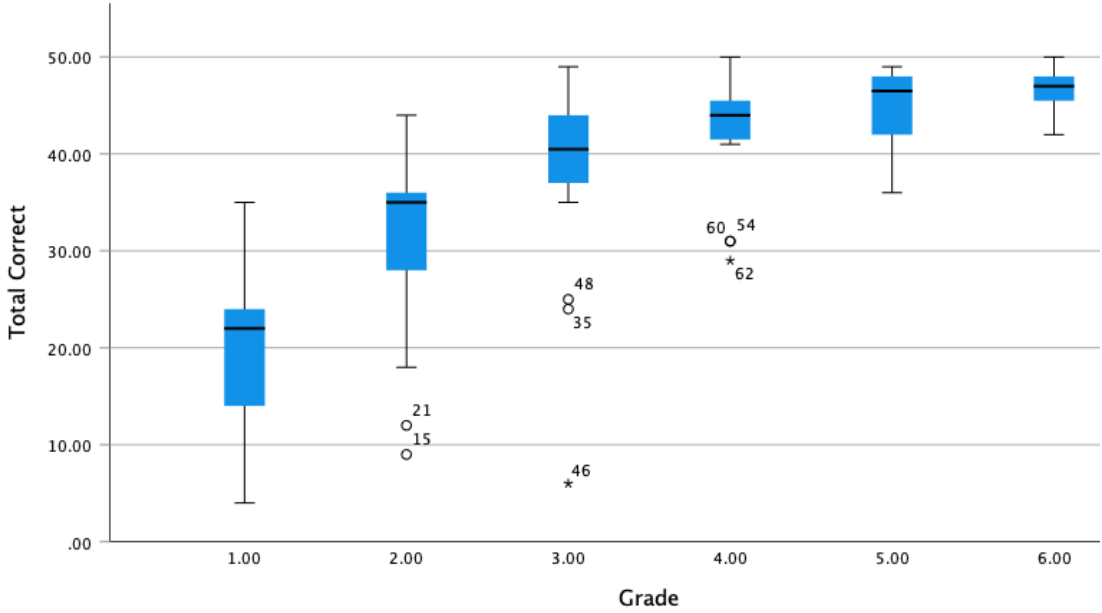
**Figure 1a**

*Distribution of correct items on the UPT-2 across the entire sample.*



**Figure 1b**

*Distribution of correct items on the UPT-2 by grade.*

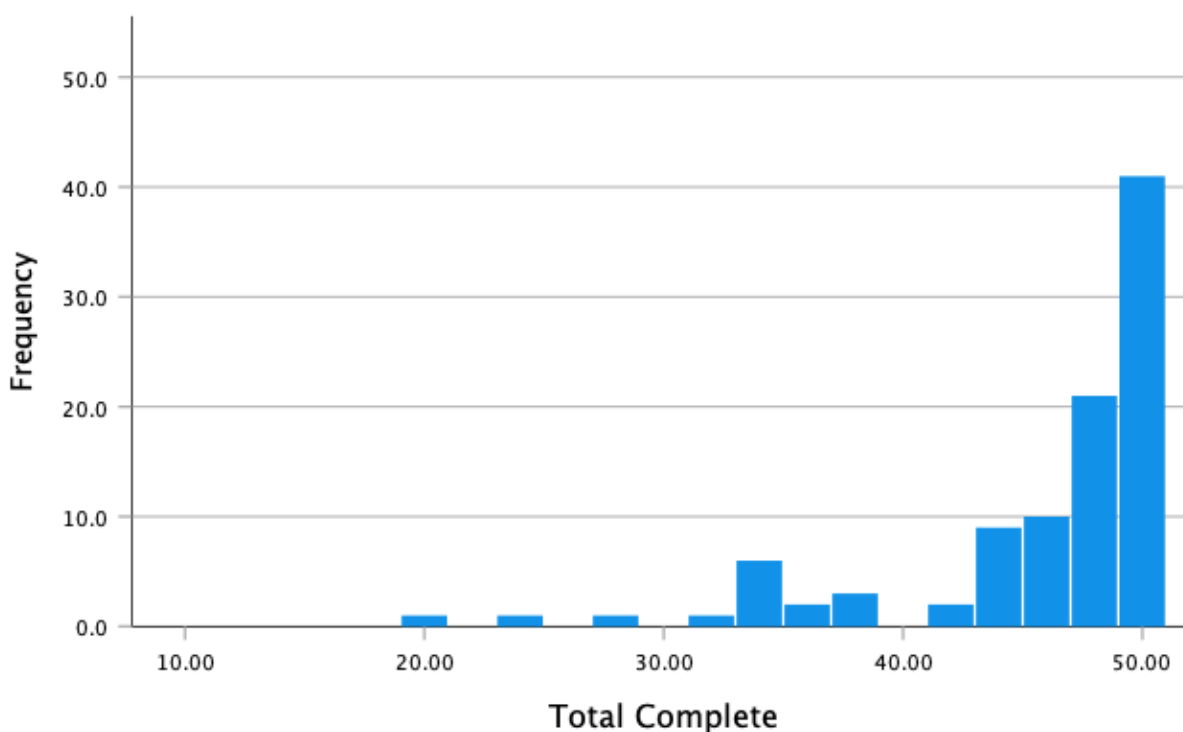


### ***UPT-2 Total Complete Items***

Across the entire sample, participants completed an average of 45.38 ( $SD=6.19$ ) items out of 50 on the UPT. The number of items completed also increased developmentally, with a significant effect of grade on total complete items on the UPT-2 [ $\chi^2(90)=115.57, p<.05$ ]. There was no significant difference between males ( $M=44.84, SD=5.75$ ) and females ( $M=45.96, SD=6.64$ ) on UPT-2 total complete scores [ $W=2539, p=.13$ ]. Cronbach's alpha for the total items complete on the UPT-2 revealed good internal consistency ( $\alpha=.91$ ). See Figure 2a and 2b for distribution of complete items across the entire sample and by grade.

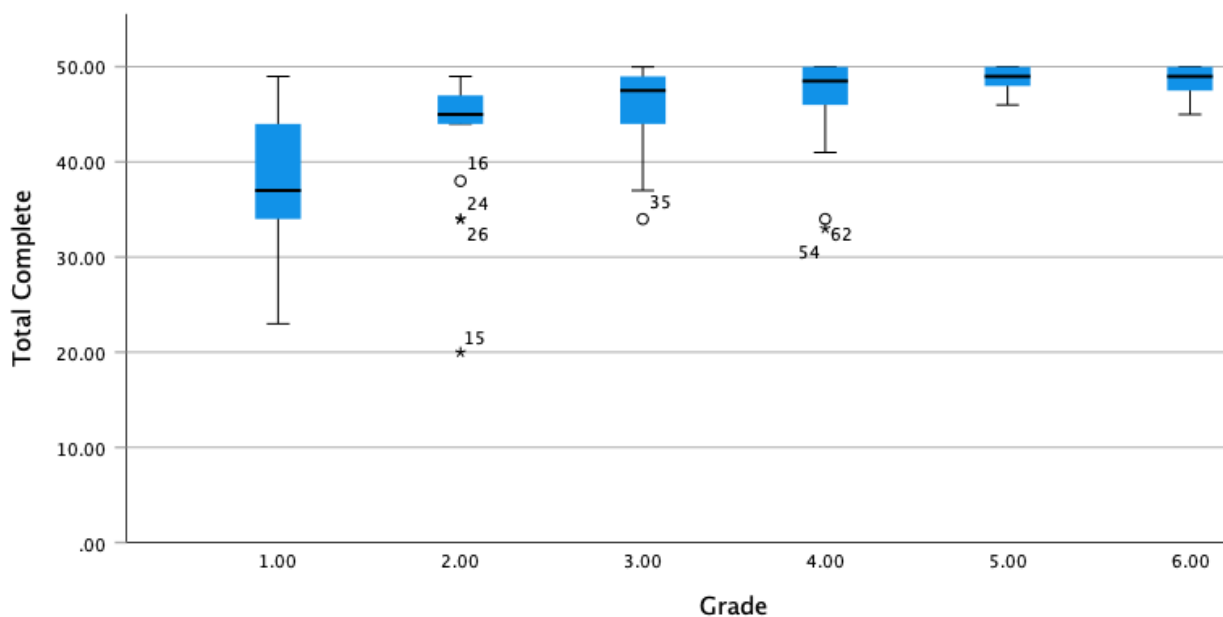
#### **Figure 2a**

*Distribution of complete items on the UPT-2 across the entire sample.*



**Figure 2b**

*Distribution of complete items on the UPT-2 by grade.*

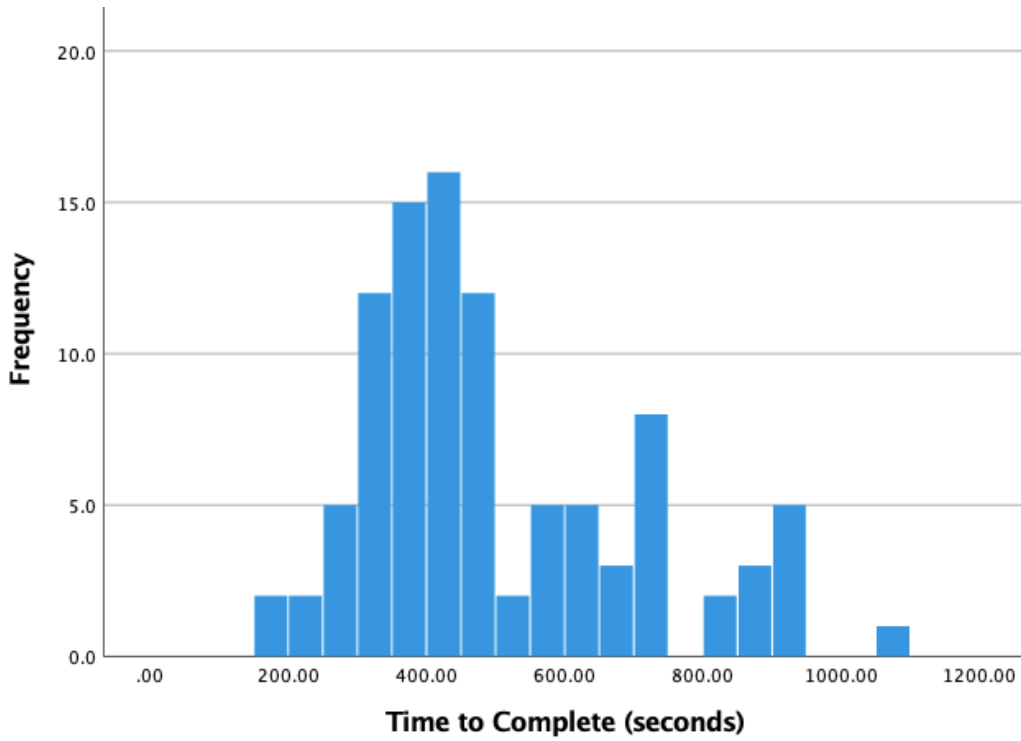


### ***UPT-2 Time to Complete***

It took participants, on average, 8 minutes and 18 seconds ( $M=498.01$  seconds,  $SD=194.07$ ) to complete the UPT-2. See Figure 3a. The amount of time it took to complete the task decreased developmentally, with a significant effect of grade on time to complete items on the UPT-2 [ $\chi^2(5)=12.09$ ,  $p<.05$ ]. See Figure 3b. There was no significant difference between males ( $M=502.88$ ,  $SD=203.89$ ) and females ( $M=492.72$ ,  $SD=184.88$ ) on UPT-2 time taken to complete the UPT-2 [ $W=2327$ ,  $p=.997$ ].

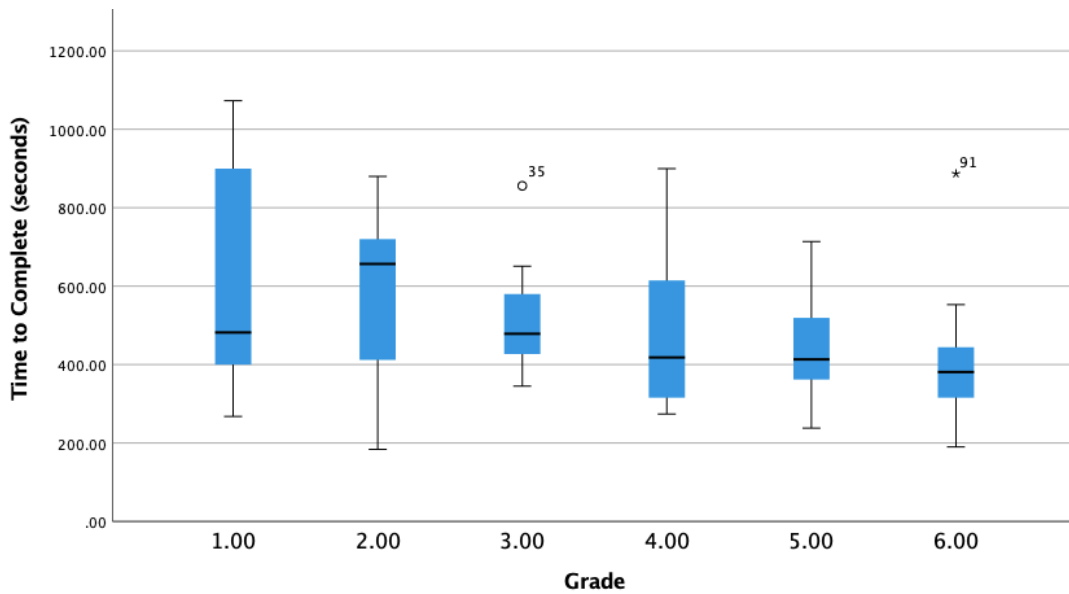
**Figure 3a**

*Distribution of Time to Complete the UPT-2 across the entire sample.*



**Figure 3b**

*Distribution of Time to Complete the UPT-2 (in seconds) across the entire sample*



### Descriptive Statistics and Scoring on the Metacognitive Ratings of Effort

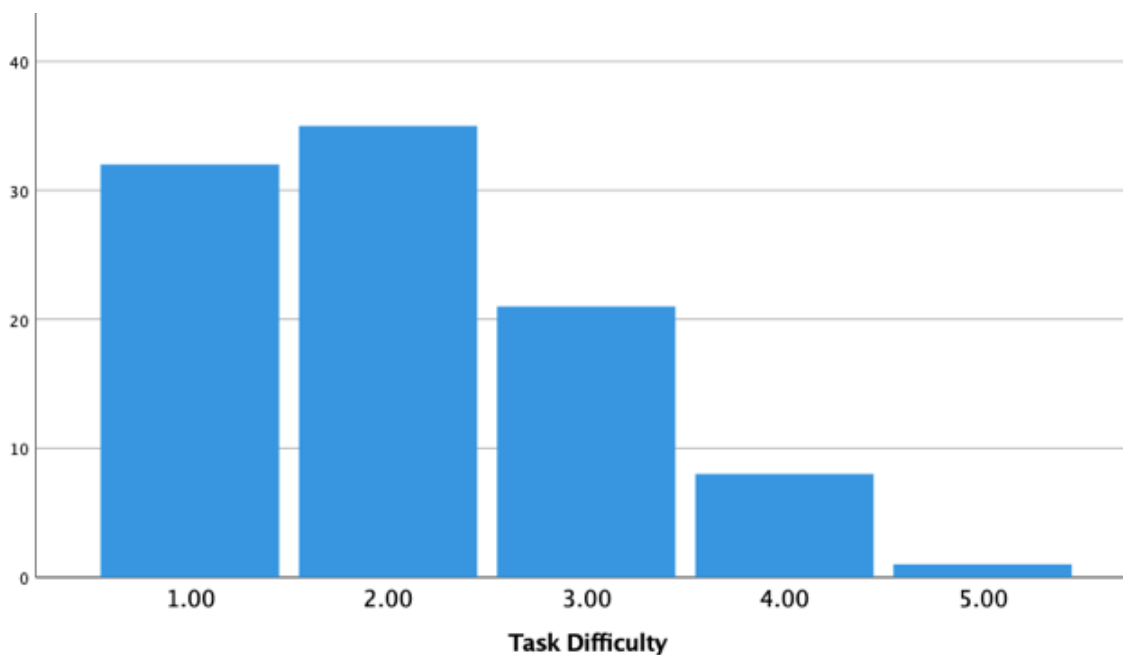
Participants rated their effort on the UPT-2 according to four questions (“How difficult was the task?”, “How hard did you try?”, “How much brainpower did you use?”, and “How did using brainpower make you feel?”). See Table 2.

#### *Task Difficulty*

There was clear evidence of a moderate, positive skew on the rating of task difficulty. See Figure 4. According to a 5-point Likert scale with 5 indicating the greatest difficulty, the average task difficulty rating was 2.08 ( $SD=.99$ ). This indicates that the UPT-2 was perceived as relatively easy. There were no gender differences on ratings of task difficulty.

#### **Figure 4**

*Distribution of Task Difficulty across the entire sample.*

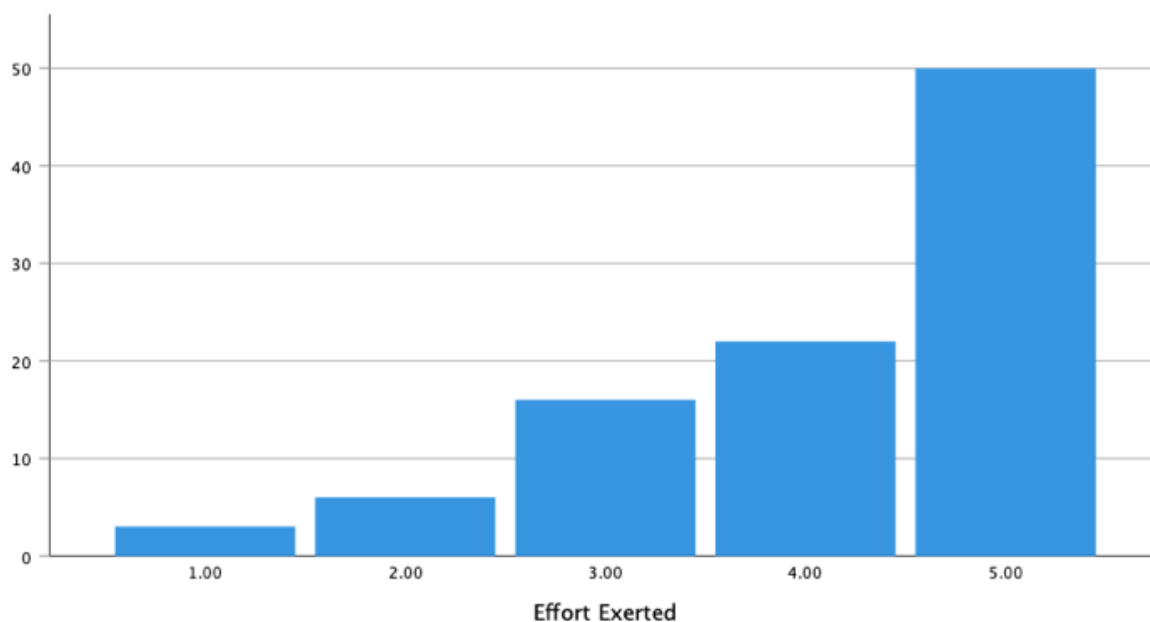


#### *Effort Exerted*

There was clear evidence of a highly negative skew on the rating of effort exerted. See Figure 5. According to a 5-point Likert scale with 5 indicating greater effort exerted, the average effort exerted on the UPT-2 was 4.13 ( $SD=1.10$ ). The higher average of effort exerted indicates that participants tried their best while completing the UPT-2. Males ( $M=3.78$ ,  $SD=1.2$ ) rated effort exerted significantly lower than females ( $M=4.51$ ,  $SD=.83$ ), [ $W= 2721$ ,  $p=.001$ ].

### Figure 5

*Distribution of Effort Exerted across entire sample.*



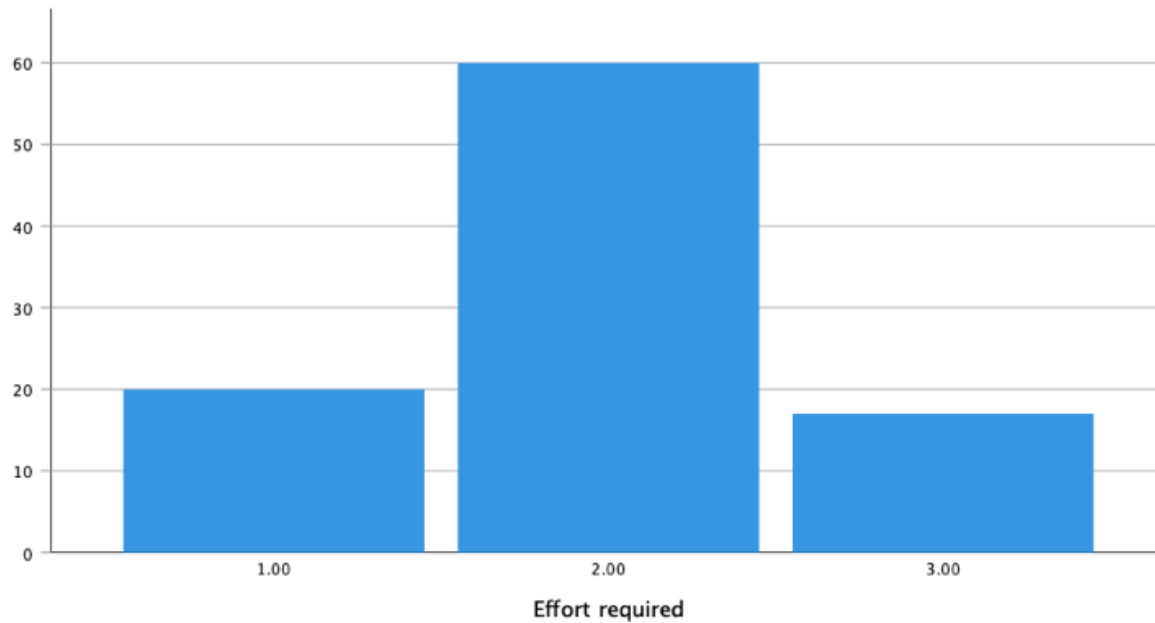
### *Effort Required*

On a 3-point Likert scale with 3 indicating greater effort required, participants reported an average rating of 1.97 ( $SD=.62$ ), indicating a need to use more brainpower to complete the UPT-2. This unimodal distribution highlights the tedious nature of the UPT-2 requiring consistent brainpower. See Figure 6. Males ( $M=1.76$ ,  $SD=.56$ ) rated effort required significantly lower than females ( $M=2.19$ ,  $SD=.61$ ), [ $W= 2712.5$ ,  $p<.001$ ].

### Figure 6



*Distribution of Effort Required across the entire sample.*

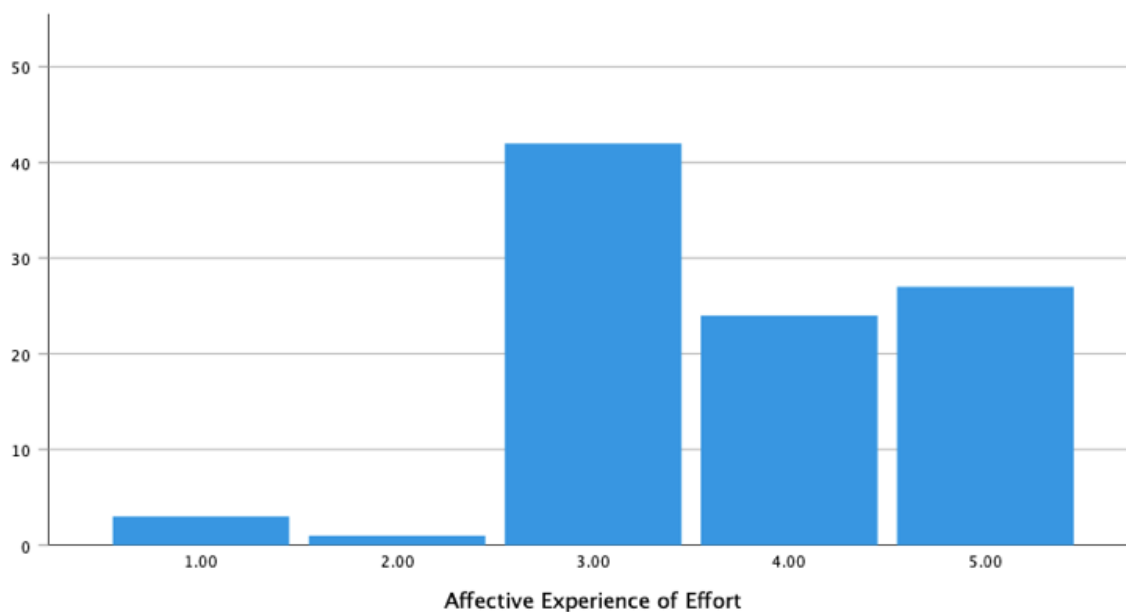


### *Affective Experience of Effort*

According to a 5-point Likert scale with 1 indicating that using brainpower made them feel negatively, the average affective experience of effort was 3.73 ( $SD=.98$ ). This multimodal distribution suggests that participants felt more positive than negative following an effortful experience. See Figure 7. Males ( $M=3.44$ ,  $SD=.86$ ) rated their affective experience of effort significantly lower than females ( $M=4.04$ ,  $SD=1.02$ ), [ $W= 2758$ ,  $p<.001$ ].

**Figure 7**

*Distribution of Affective Experience of Effort across entire sample.*



### **Correlations Between Age and UPT-2 Performance, UPT-2 Metacognitive Ratings**

Correlations of age with all variables were calculated using Spearman correlations. Age was significantly related to UPT-2 total correct ( $r_s = .78$ ,  $p < .01$ ), total complete ( $r_s = .60$ ,  $p < .01$ ), and composite performance across all domains (math, language, symbol). Results indicated that time taken to complete the UPT-2 was negatively correlated with age ( $r_s = -.36$ ,  $p < .01$ ). Age was significantly related to student's ratings of task difficulty ( $r_s = -.46$ ,  $p < .01$ ) and effort required ( $r_s = -.44$ ,  $p < .01$ ), but not with effort exerted or affective experience of effort. That is, older participants rated the UPT-2 as less difficult and requiring less effort than younger participants.

Parent-rated executive function on the BDEFS-CA did not significantly correlate with age. Performance on the Stroop Test ( $r_s = -.58$ ,  $p < .01$ ) and Trail Making Test ( $r_s = -.62$ ,  $p < .01$ )

was correlated with age. Age and academic achievement were significantly correlated, including math ( $r_s = .80, p < .01$ ) and reading abilities ( $r_s = .75, p < .01$ ). See Table 3 for age correlations.

**Table 3**

*Spearman correlations between all variables and age*

	Age	N
<b><u>UPT-2 Performance and Ratings</u></b>		
UPT-2 correct responses – all items	.78**	98
UPT-2 correct responses – language items	.71**	97
UPT-2 correct responses – math items	.75**	97
UPT-2 correct responses – symbol items	.60*	98
UPT-2 completed responses – all items	.60**	98
UPT-2 completed responses – language items	.50**	98
UPT-2 completed responses – math items	.58**	98
UPT-2 completed responses – symbol items	.30**	98
UPT-2 composite performance – all items	.77**	98
UPT-2 composite performance – language items	.71**	97
UPT-2 composite performance – math items	.74**	97
UPT-2 composite performance – symbol items	.57**	98
UPT-2 Total Circled (Scored complete but not accurate)	-.66**	98
UPT-2 Total Missed (Scored incomplete and not accurate)	-.59**	98
UPT-2 Completion Time	-.36**	98
Task difficulty rating	-.46**	97
Effort exerted rating	-.02	97
Effort required rating	-.44**	97
Affective experience of effort rating	-.10	97
<b><u>Executive Function Ratings and Performance-Based Tasks</u></b>		
BDEFS-CA <sup>1</sup> Parent Rating	.08	67
Stroop Interference Time <sup>1</sup>	-.58**	98
Trail-Making Part B minus Part A Time <sup>1</sup>	-.62**	94
<b><u>Academic Tasks</u></b>		
TOWRE-2 <sup>2</sup> Phonemic Decoding Efficiency raw score	.70**	98
TOWRE-2 Sight Word Efficiency raw score	.76**	98
TOWRE-2 Composite z-score	.75**	98
WJ-IV Math Calculation Raw Score	.77**	97
WJ-IV Math Fluency Raw Score	.79**	96
WJ-IV Math Composite z-score	.80**	96

*Note.* \*  $p < .05$ ; \*\*  $p < .01$ . UPT-2=Unstructured Performance Task Version 2, WJTA=Woodcock-Johnson Tests of Achievement IV, TOWRE-2=Test of Word Reading Efficiency, BDEFS-CA=Barkley Deficits in Executive Functioning Scale.

<sup>1</sup> A higher score indicates more EF difficulties.

<sup>2</sup> A higher score indicates better performance

## **Correlates of the Unstructured Performance Task - 2**

### ***Performance-based Tasks and Rating Scale of Executive Function***

Correlations of the UPT-2 with the BDEFS-CA, Stroop Test, and Trail Making Test were explored using Spearman correlations (see Table 4). The Stroop Test was significantly related to total correct items ( $r_s = -.56, p < .001$ ), total complete items ( $r_s = -.48, p < .001$ ), and composite performance ( $r_s = -.56, p < .001$ ) on the UPT-2. These correlations remained consistent across the math, language, and symbol domains of the UPT-2. The Trail Making Test was also significantly related to total correct items ( $r_s = -.64, p < .001$ ), total complete items ( $r_s = -.50, p < .001$ ), and composite performance ( $r_s = -.63, p < .001$ ) on the UPT-2, and this held true across all three domains. See Figure 8a for a correlation distribution of UPT-2 composite performance and combined z-score of performance-based tasks (Stroop and Trail Making Test). Across all domains, the UPT-2 total correct, total complete, and composite performance was not significantly related to the BDEFS-CA parent rating scale. This is consistent with Wanstall's (2019) findings on the UPT. See Figure 8b for a correlation distribution of UPT-2 composite performance and the BDEFS-CA. As expected, the BDEFS-CA was not significantly related to Stroop Test or the Trail Making Test. Future UPT-2 studies should consider the developmental sensitivity of these tasks.

**Table 4**

*Spearman Correlations between UPT-2 variables, Performance-based EF Tasks, and a Rating Scale of EF*

	<u>Performance-Based EF Tasks</u>			<u>Rating Scale of EF</u>
	Stroop <sup>1</sup> Total Score	Trails <sup>1</sup> Total Score	Combined <sup>1</sup> z-score	BDEFS-CA <sup>1</sup> Total Score
<b><u>UPT-2 Correct</u></b> <sup>2</sup>				
Total	-.56**	-.64**	-.69**	.06
Math	-.55**	-.54**	-.62**	.10
Language	-.56**	-.65**	-.69**	.05
Symbol	-.36**	-.49**	-.46**	.05
<b><u>UPT-2 Complete</u></b> <sup>2</sup>				
Total	-.48**	-.50**	-.57**	-.01
Math	-.46**	-.45**	-.52**	-.08
Language	-.46**	-.48**	-.54**	-.02
Symbol	-.22*	-.24*	-.24*	
<b><u>UPT-2 Composite</u></b> <sup>2</sup>				
Total	-.56**	-.63**	-.69**	.06
Math	-.55**	-.55**	-.64**	.08
Language	-.56**	-.63**	-.70**	.06
Symbol	-.33**	-.44**	-.45**	.04

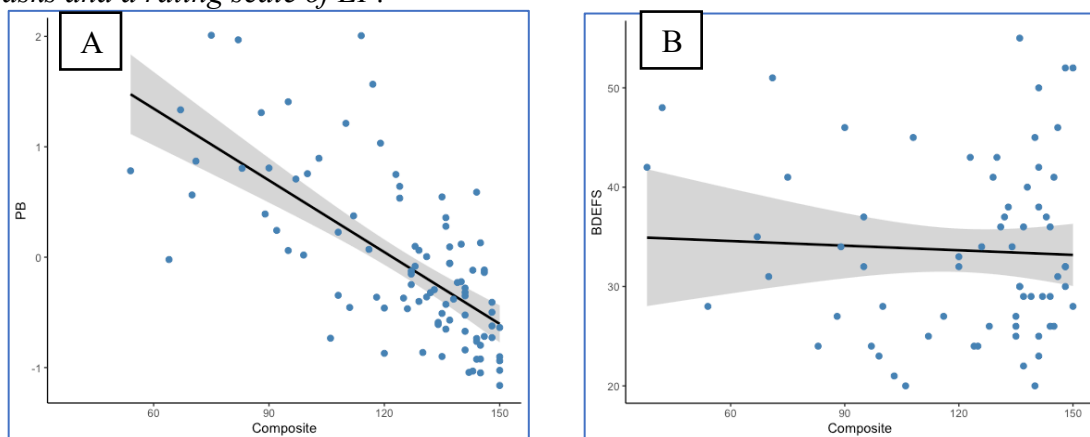
*Note.* \*  $p < .05$ ; \*\*  $p < .01$ . UPT-2=Unstructured Performance Task Version 2, BDEFS-CA = Barkley Deficits in Executive Functioning Scale.

<sup>1</sup> A higher score indicates more EF difficulties.

<sup>2</sup> A higher score indicates better performance

**Figure 8**

Scatter plots of correlations between composite performance on the UPT-2 and performance-based tasks and a rating scale of EF.



*Note.* The black lines represent the linear fit and the shaded areas represent the 95% confidence regions. Composite performance on the UPT-2 are shown on the x-axes and outcome measures on the y-axes, including Performance-based Tasks of EF (A), Rating Scale of EF (B).

### ***Academic Abilities***

Math abilities were significantly related to total correct items ( $r_s=.79$ ,  $p<.001$ ), total complete items ( $r_s=.57$ ,  $p<.001$ ), and composite performance ( $r_s=.75$ ,  $p<.001$ ) on the UPT-2. These correlations remained consistent across the math, language, and symbol domains of the UPT-2. Reading abilities were also significantly related to total correct items ( $r_s=.67$ ,  $p<.001$ ), total complete items ( $r_s=.47$ ,  $p<.001$ ), and composite performance ( $r_s=.65$ ,  $p<.001$ ) on the UPT-2, and this held true across domains of the UPT-2. This replicates findings from Wanstall (2019) on the UPT, suggesting that both academic skills and performance on the UPT-2 increase with age. See Table 5 and Figure 9a-d for correlations.

**Table 5***Spearman Correlations between UPT-2 variables and Academic Abilities*

	<u>Math Abilities</u>			<u>Reading Abilities</u>		
	WJ-IV Calculation	WJ-IV Fluency	Combined z-score	TOWRE- 2 PD Raw Score	TOWRE-2 SWR Raw Score	Combined z-score
<b><u>UPT-2 Correct</u><sup>2</sup></b>						
Total	.70**	.67**	.79**	.65**	.65**	.67**
Math	.67**	.63**	.75**	.63**	.61**	.64**
Language	.65**	.60**	.71**	.69**	.70**	.72**
Symbol	.55**	.58**	.66**	.38**	.41**	.41**
<b><u>UPT-2 Complete</u><sup>2</sup></b>						
Total	.53**	.47**	.57**	.45**	.46**	.47**
Math	.45**	.41**	.51**	.42**	.42**	.43**
Language	.54**	.45**	.56**	.51**	.53**	.54**
Symbol	.28**	.34**	.36**	.05	.14	.09
<b><u>UPT-2 Composite</u><sup>2</sup></b>						
Total	.72**	.72**	.75**	.63**	.63**	.65**
Math	.67**	.68**	.70**	.62**	.60**	.63**
Language	.68**	.65**	.68**	.69**	.69**	.71**
Symbol	.54**	.58**	.60**	.33**	.37**	.36**

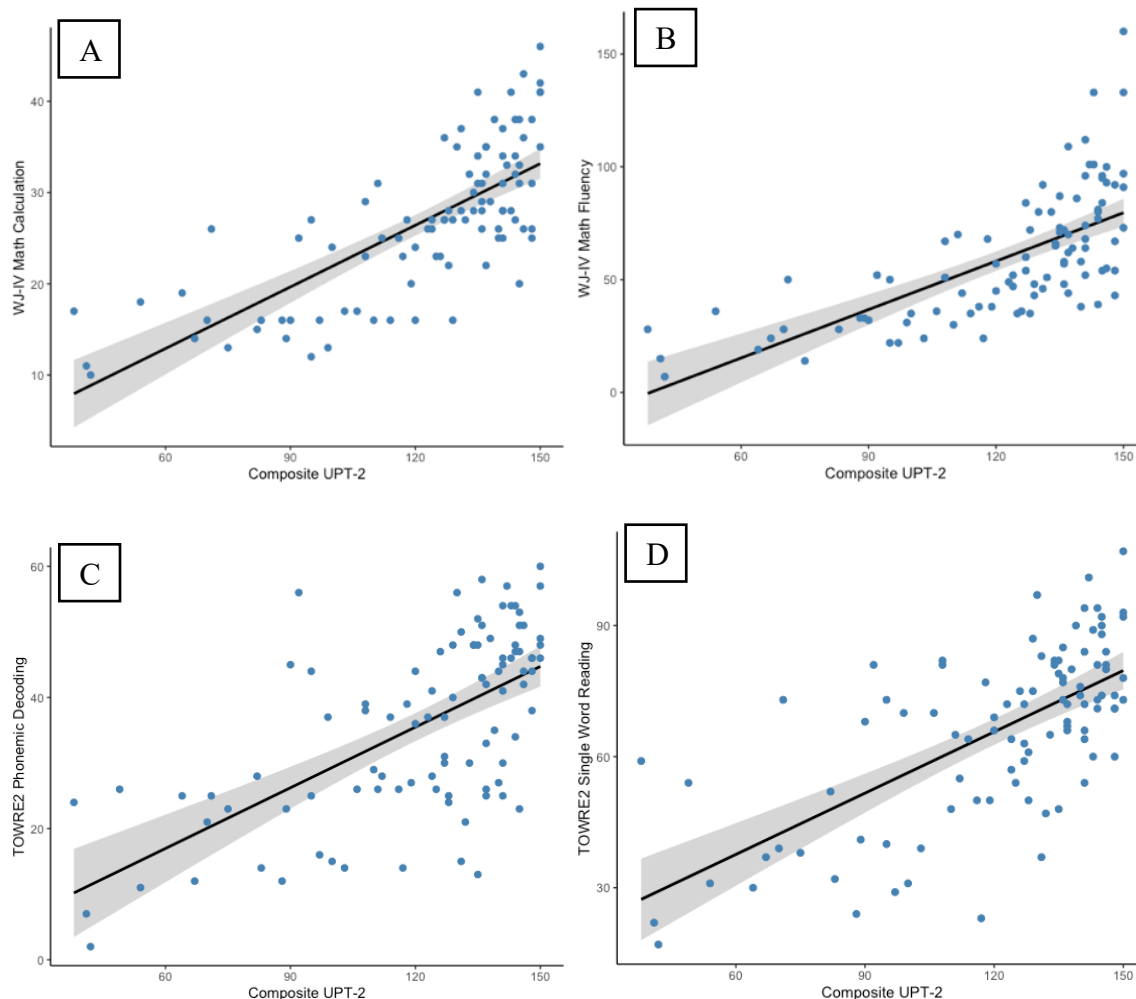
*Note.* \*  $p < .05$ ; \*\*  $p < .01$ . UPT-2=Unstructured Performance Task Version 2, WJ-IV=Woodcock-Johnson Tests of Achievement – Fourth Edition, TOWRE-2=Test of Word Reading Efficiency – Second Edition,

<sup>1</sup> A higher score indicates more EF difficulties.

<sup>2</sup> A higher score indicates better performance

**Figure 9**

Scatter plots of correlations between composite performance on the UPT-2 and measures of academic achievement (WJTA and TOWRE).



*Note.* The black lines represent the linear fit and the shaded areas represent the 95% confidence regions. Composite performance 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).

### ***Metacognitive Ratings of Effort***

Task difficulty was significantly related to total correct items ( $r_s = -.35$ ,  $p < .001$ ), total complete items ( $r_s = -.23$ ,  $p < .001$ ), and composite performance ( $r_s = -.32$ ,  $p < .001$ ) on the UPT-2.

These correlations remained consistent across the math, language, and symbol domains of the



UPT-2. Effort exerted was not significantly related to total correct items, total complete items, or composite performance on the UPT-2 across any of the three domains. Effort required was significantly related to total correct items ( $r_s = -.33, p < .001$ ), total complete items ( $r_s = -.21, p < .001$ ), and composite performance ( $r_s = -.30, p < .001$ ) on the UPT-2, and this held true across domains of the UPT-2. The affective experience of effort was not significantly related to total correct items, total complete items, or composite performance on the UPT-2 across any of the three domains. See Table 6 for correlations.

**Table 6**

*Spearman Correlations between UPT-2 variables and Metacognitive Ratings of Effort*

	Task Difficulty	Effort Exerted	Effort Required	Affective Experience of Effort
<b><u>UPT-2 Correct</u></b>				
Total	-.35**	<.01	-.33**	-.04
Math	-.34**	-.08	-.32**	-.05
Language	-.26*	.03	-.28**	-.02
Symbol	-.25*	.10	-.26*	-.03
<b><u>UPT-2 Complete</u></b>				
Total	-.23*	<.01	-.21*	<.01
Math	-.16	-.10	-.21*	-.07
Language	-.20	.03	-.21*	<.01
Symbol	-.22*	.16	-.09	-.10
<b><u>UPT-2 Composite</u></b>				
Total	-.32**	<.01	-.30**	<.05
Math	-.31**	-.10	-.29**	-.07
Language	-.24**	<.05	-.26**	-.02
Symbol	-.26**	.15	-.21*	.03

Note. \*  $p < .05$ ; \*\*  $p < .01$ . UPT-2=Unstructured Performance Task Version 2

**Correlates within Metacognitive Ratings of Effort**

Task difficulty was significantly related to effort exerted ( $r_s = .28, p < .01$ ) and effort required ( $r_s = .55, p < .01$ ), but not affective experience of effort ( $r_s = .07, p = .47$ ). Effort exerted was significantly related to effort required ( $r_s = .32, p < .01$ ) and the affective experience of effort

( $r_s=.35$ ,  $p < .01$ ). Effort required was not significantly related to the affective experience of effort ( $r_s=.15$ ,  $p = .13$ ). See Table 7 for correlations.

**Table 7**

*Spearman correlations within metacognitive ratings of effort (task difficulty, effort exerted, effort required, and affective experience of effort)*

	Task difficulty	Effort exerted	Effort required
Task difficulty	-	-	-
Effort exerted	.28**	-	-
Effort required	.55**	.32**	-
Affective experience of effort	.07	.35**	.15

Note. \*\*  $p < .01$

A median split of UPT-2 performance was used to further explore the association between metacognitive ratings of effort and performance on the UPT-2. In participants who performed “low”, task difficulty was significantly related to effort exerted ( $r_s=.40$ ,  $p < .01$ ) and effort required ( $r_s=.58$ ,  $p < .01$ ), but not affective experience of effort ( $r_s=-.005$ ,  $p = .97$ ). Effort exerted was significantly related to effort required ( $r_s=.33$ ,  $p < .01$ ), but not affective experience of effort ( $r_s=.25$ ,  $p < .08$ ). Effort required was not significantly related to the affective experience of effort ( $r_s=.12$ ,  $p = .43$ ). The UPT-2 composite performance did not correlate with any metacognitive ratings of effort in the “low” category. See Table 8 for correlations.

In participants who performed “high”, task difficulty was significantly related to effort required ( $r_s=.39$ ,  $p < .01$ ), but not effort exerted ( $r_s=-.20$ ,  $p = .16$ ) or affective experience of effort ( $r_s=-.11$ ,  $p = .45$ ). Effort exerted was significantly related to effort required ( $r_s=.40$ ,  $p < .01$ ) and affective experience of effort ( $r_s=.48$ ,  $p < .01$ ). Effort required was not significantly related to the affective experience of effort ( $r_s=.14$ ,  $p = .34$ ). The UPT-2 composite performance was significantly related to task difficulty ( $r_s=-.33$ ,  $p < .01$ ), but not effort exerted, effort required, or the affective experience of effort. See Table 8 for correlations.

**Table 8**

*Spearman correlations within metacognitive ratings of effort (task difficulty, effort exerted, effort required, and affective experience of effort) and UPT-2 composite performance in low vs. high scorers.*

	Effort exerted	Effort required	Affective Experience	UPT-2 Composite
<u>N= 47 (UPT-2 Low Scorers: Composite range=38-130)</u>				
Task difficulty	.40**	.58**	-.005	.03
Effort exerted		.33**	.25	.04
Effort required			.12	-.11
Affective experience of effort			-	.03
<u>N=50 (UPT-2 High Scorers: Composite range=131-150)</u>				
Task difficulty	.20	.39**	.11	-.33*
Effort exerted		.40**	.48**	-.11
Effort required			.14	-.11
Affective experience of effort				.08

## **Predictors of the Unstructured Performance Task – 2**

### ***Metacognitive Ratings***

A multiple regression was run to predict UPT-2 performance from metacognitive ratings against the predicted values. There was independence of residuals, as assessed by a Durbin-Watson statistic of 1.313. There was homoscedasticity, as assessed by visual inspection of a plot of studentized residuals versus unstandardized predicted values. There was no evidence of multicollinearity, as assessed by tolerance values greater than 0.1. There were no studentized deleted residuals greater than  $\pm 3$  standard deviations, no leverage values greater than 0.2, and values for Cook's distance above 1. The assumption of normality was met, as assessed by a Q-Q Plot. The multiple regression model statistically significantly predicted UPT-2 composite performance,  $F(4, 91) = 3.111, p < .05, \text{adj. } R^2 = .081$ . Regression coefficients and standard errors can be found in Table 9.

**Table 9**

*Regression results predicting UPT-2 Composite Performance from metacognitive ratings of effort*

UPT-2 Composite Performance	<i>B</i>	95% CI for <i>B</i>		<i>SE(B)</i>	$\beta$	$R^2$
		<i>LL</i>	<i>UL</i>			
Model						.119***
Task difficulty	-3.18	-9.91	3.55	3.39	-.11	
Effort exerted	4.44	-1.26	10.13	2.87	.17	
Effort required	-13.43	-24.38	-2.47	5.51	-.30*	
Affective experience of effort	1.71	-4.10	7.53	2.93	.06	

*Note.* Model = “Enter” method in SPSS Statistics; *B* = unstandardized regression coefficient; CI = confidence interval; *LL* = lower limit; *UL* = upper limit; *SE B* = standard error of the coefficient;  $\beta$  = standardized coefficient;  $R^2$  = coefficient of determination;  $\Delta R^2$  = adjusted  $R^2$ . \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

***Performance-based Tasks of Executive Function and Rating of Effort Required, after controlling for Age***

A hierarchical multiple regression was conducted to determine if the addition of EF tasks and the effort required rating improved the prediction of UPT-2 composite performance over and above age alone. See Table 10 for full details on each regression model. The full model of age, EF tasks, and the effort required rating to predict UPT-2 composite performance (Model 2) was statistically significant,  $R^2 = .62$ ,  $F(3, 89) = 47.81$ ,  $p < .001$ ; adjusted  $R^2 = .60$ . The addition of EF tasks to the prediction of UPT-2 composite performance (Model 2) led to a statistically significant increase in  $R^2$  of .06,  $F(2, 89) = 6.95$ ,  $p < .01$ . This finding suggests that younger children and those with weaker EF (according to the EF composite) performed worse on the UPT-2. Regression coefficients and standard errors can be found in Table 10.

**Table 10***Hierarchical Regression Analysis for UPT-2 Composite Performance (N=98)*

Step	Predictor	$\Delta R^2$	F Change	B	SE B	$\beta$
1	Age	.55	114.54***	.85	.08	.75***
2	Age	.60	6.95***	.62	.11	.55***
	EF Composite			-10.68	2.89	-.34***
	Effort Required			3.30	2.84	.09

Note. Model = "Enter" method in SPSS Statistics; *B* = unstandardized regression coefficient; *SE B* = standard error of the coefficient;  $\beta$  = standardized coefficient;  $\Delta R^2$  = adjusted  $R^2$ .

\*\*\* $p < .001$ .

***Behavioral Rating scale of Executive Function and Effort Required Rating***

A multiple regression was run to predict UPT-2 performance from a rating scale of EF and the effort required rating. Given that age was not associated with the BDEFS-CA, it was not included in the model. There was independence of residuals, as assessed by a Durbin-Watson statistic of 2.56. There was homoscedasticity, as assessed by visual inspection of a plot of studentized residuals versus unstandardized predicted values. There was no evidence of multicollinearity, as assessed by tolerance values greater than 0.1. There were no studentized deleted residuals greater than  $\pm 3$  standard deviations, no leverage values greater than 0.2, and values for Cook's distance above 1. The assumption of normality was met, as assessed by a Q-Q Plot. The multiple regression model statistically significantly predicted UPT-2 composite performance,  $F(4, 91) = 3.111, p < .05, \text{adj. } R^2 = .081$ . Regression coefficients and standard errors can be found in Table 11.

**Table 11**

*Regression results predicting UPT-2 Composite Performance from an EF rating scale and effort required rating*

UPT-2 Composite Performance	<i>B</i>	95% CI for <i>B</i>		SE( <i>B</i> )	$\beta$	$R^2$
		<i>LL</i>	<i>UL</i>			
Model						.08
BDEFS	-.31	-1.07	.46	.39	-.10	
Effort required	-13.53	-25.34	-1.73	5.91	-.28*	

*Note.* Model = “Enter” method in SPSS Statistics; *B* = unstandardized regression coefficient; CI = confidence interval; *LL* = lower limit; *UL* = upper limit; *SE B* = standard error of the coefficient;  $\beta$  = standardized coefficient;  $R^2$  = coefficient of determination

\* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

### **Predictors of Reading Abilities**

#### *Age, UPT-2, Effort Required, EF tasks, and an EF rating scale*

A hierarchical multiple regression was run to determine if the addition of EF tasks and an EF rating scale improved the prediction of Reading Abilities over and above age, UPT-2 composite performance, and the effort required rating. See Table 12 for full details on each regression model. The full model of age, UPT-2, effort required, EF tasks, and an EF rating scale to predict Reading Abilities (Model 2) was statistically significant,  $R^2 = .65$ ,  $F(5, 59) = 21.73$ ,  $p < .001$ ; adjusted  $R^2 = .62$ . The addition of EF tasks and an EF rating scale to the prediction of Reading Abilities (Model 2) led to a statistically significant increase in  $R^2$  of .07,  $F(2, 59) = 6.10$ ,  $p < .01$ . Regression coefficients and standard errors can be found in Table 12.

**Table 12***Hierarchical Regression Analysis for Reading Abilities (N=98)*

Step	Predictor	$\Delta R^2$	F Change	B	SE B	$\beta$
1		.55	27.54***			
	Age			.02	.01	.53***
	UPT-2			.01	.01	.21
	Effort Required			-.21	.15	-.13
2		.62	6.10**			
	Age			.02	.01	.34*
	UPT-2			.002	.01	.06
	Effort Required			-.12	.14	-.07
	EF Tasks			-.56	.17	-.41***
	BDEFS			.02	.01	.14

*Note.* Model = “Enter” method in SPSS Statistics; *B* = unstandardized regression coefficient; *SE B* = standard error of the coefficient;  $\beta$  = standardized coefficient;  $\Delta R^2$  = adjusted  $R^2$ .

\*\*\* $p < .001$ .

### **Predictors of Math Abilities**

#### ***Age, UPT-2, Effort Required, EF tasks, and an EF rating scale***

A hierarchical multiple regression was run to determine if the addition of EF tasks and an EF rating scale improved the prediction of Math Abilities over and above age, UPT-2 composite performance, and the effort required rating. See Table 13 for full details on each regression model. The full model of age, UPT-2, effort required, EF tasks, and an EF rating scale to predict Math Abilities (Model 2) was statistically significant,  $R^2 = .68$ ,  $F(5, 59) = 24.71$ ,  $p < .001$ ; adjusted  $R^2 = .65$ . The addition of EF tasks and an EF rating scale to the prediction of Math Abilities (Model 2) did not lead to a statistically significant increase in  $R^2$ . ,  $F(2, 59) = .84$ ,  $p = .435$ . Regression coefficients and standard errors can be found in Table 13.

**Table 13***Hierarchical Regression Analysis for Math Abilities (N=98)*

Step	Predictor	$\Delta R^2$	F Change	B	SE B	$\beta$
1		.65	40.83***			
	Age			.02	.01	.45***
	UPT-2			.01	.01	.32**
	Effort Required			-.33	.12	-.21**
2		.65	.843			
	Age			.02	.01	.50***
	UPT-2			.01	.01	.30*
	Effort Required			-.35	.13	-.23**
	EF Tasks			.01	.15	.01
	BDEFS			-.01	.01	-.10

Note. Model = "Enter" method in SPSS Statistics; B = unstandardized regression coefficient; SE B = standard error of the coefficient;  $\beta$  = standardized coefficient;  $\Delta R^2$  = adjusted  $R^2$ .

\*\*\* $p < .001$ .

**Table 14. Summary of Results**

Hypothesis	Result
<p><b><u>Performance-Based Tasks and Rating Scale of EF</u></b></p> <p><b>1a)</b> UPT-2 performance would correlate with EF performance-based tasks, but not with EF ratings.</p> <p><b>1b)</b> EF tasks would not correlate with EF ratings</p>	<p><b>1a)</b> Better UPT-2 performance was correlated with better scores on performance-based tasks of EF, but not with the EF rating scale, which is consistent with our hypothesis.</p> <p><b>1b)</b> EF performance-based tasks did not correlate with EF ratings, which is consistent with our hypothesis.</p>
<p><b><u>Academic Abilities</u></b></p> <p><b>2)</b> UPT-2 performance would correlate with reading and math achievement</p>	<p><b>2)</b> Better UPT-2 performance positively correlated with better reading and math achievement, which is consistent with our hypothesis.</p>
<p><b><u>Correlations within metacognitive ratings of effort</u></b></p> <p><b>3a)</b> Task difficulty would correlate with effort exerted</p> <p><b>3b)</b> Effort exerted would correlate with effort required</p>	<p><b>3a) Ratings of</b> task difficulty positively correlate with ratings of effort exerted, which is consistent with our hypothesis.</p> <p><b>3b)</b> Higher ratings of effort exerted positively correlated with higher ratings</p>



<p><b>3c)</b> Effort exerted would correlate with the affective experience of effort</p>	<p>of effort required, which is consistent with our hypothesis  <b>3c)</b> Higher ratings of effort exerted positively correlated with higher ratings of affective experience of effort, which is consistent with our hypothesis.</p>
<p><b><u>Metacognitive Ratings of Effort and the UPT-2</u></b></p> <p><b>4a)</b> UPT-2 performance would negatively correlate with task difficulty</p> <p><b>4b)</b> UPT-2 performance would correlate with effort exerted</p> <p><b>4c)</b> UPT-2 performance would negatively correlate with effort required</p> <p><b>4d)</b> UPT-2 performance would not correlate with affective experience of effort</p>	<p><b>4a)</b> Better UPT-2 performance negatively correlated with lower ratings of task difficulty in all domains, which is consistent with our hypothesis.  <b>4b)</b> UPT-2 performance did not correlate with effort exerted in any domains, which is not consistent with our hypothesis.  <b>4c)</b> Better UPT-2 performance was negatively correlated with lower effort required ratings in all domains, which is consistent with our hypothesis.  <b>4d)</b> UPT-2 performance did not correlate with affective experience of effort in any domains, which is consistent with our hypothesis.</p>
<p><b><u>Regression Analyses</u></b></p> <p><b>5a)</b> Metacognitive ratings of effort would enter as a significant predictor of UPT-2 performance</p> <p><b>5b)</b> Performance-based EF tasks and the effort required rating would predict UPT-2 performance after controlling for age</p> <p><b>5c)</b> The rating scale of EF and the effort required rating would predict UPT-2 performance</p> <p><b>5d)</b> Neither performance-based tasks nor the rating scale of EF would emerge as significant predictors of reading abilities after controlling for age, UPT-2 performance, and effort required</p>	<p><b>5a)</b> Effort required entered as a significant predictor of UPT-2 performance, which is consistent with our hypothesis.  <b>5b)</b> Performance-based EF tasks entered as a significant predictor of UPT-2 performance after controlling for age, which is consistent with our hypothesis.  <b>5c)</b> The effort required rating entered as a significant predictor of UPT-2 performance, which is consistent with our hypothesis.  <b>5d)</b> Performance-based tasks of EF entered as a significant predictor of reading abilities after controlling for age, UPT-2 performance, and effort required, which is not consistent with our hypothesis.</p>

<p><b>5e)</b> The rating scale of EF would emerge as a significant predictor of math abilities after controlling for age, UPT-2 performance, and effort required</p>	<p><b>5e)</b> Neither performance-based tasks of EF nor the rating scale of EF entered as significant predictors of math abilities after controlling for age, UPT-2 performance, and effort required, which is not consistent with our hypothesis.</p>
--	--

### Discussion

The overarching purpose of this study was to examine metacognitive correlates of effort on the Unstructured Performance-Based Task (UPT-2). This study had 3 objectives. First, to replicate patterns of association between performance-based measures, a rating scale, and the UPT-2, as well as the associations between the UPT-2 and age found in Ledochowski et al. (2019) and Wanstall's (2019). Second, to explore the value of metacognitive ratings of effort, as well as their association with UPT-2 performance, age, and gender. Third, to further Wanstall et al.'s conceptualization of the UPT-2 as a predictor of academic ability, with the novel contribution of a metacognitive rating of effort. Results showed that there were significant associations among the UPT-2 performance, performance-based measures, behavioral rating scale of EF, and age. Hierarchical regression analyses showed that performance-based measures and behavioral rating scales of EF significantly predicted UPT-2 performance after controlling for age. There were significant associations within metacognitive ratings of effort, as well as age differences. Finally, task difficulty and effort required emerged as significant predictors of UPT-2 performance when evaluated individually, but only effort required significantly predicted UPT-2 performance when all four effort ratings were included in the regression model. Given the significance of the effort required rating in its ability to predict UPT-2 performance, we continued a series of regressions to determine the strength of its predictive ability among a behavioral rating scale of EF, performance-based tasks of EF, and age.

### **Associations Among the UPT-2, Performance-Based Measures of EF, and Behavioral Rating Scale of EF**

Several significant associations emerged among the UPT-2 variables and performance-based measures of EF. First, consistent with past literature, associations between performance-based measures and the behavioral rating scale of EF were not significant (Bodnar et al., 2007; Gray et al., 2015; Mahone et al., 2002; McAuley et al., 2010; Toplak et al., 2013). Performance on the UPT-2 across all domains (total, language, math, symbol), specifically UPT-2 total correct, UPT-2 total complete, UPT-2 composite performance, and UPT-2 completion time was associated with performance-based tasks of EF. The pattern of association among the UPT-2 and performance-based measures suggests the UPT-2 may be tapping into common factors of EF shared with performance-based measures. The strength of the correlations ranged from small to large with the smallest correlation being between UPT-2 completion time ( $r=.22$ ) and the strongest correlation being between the UPT-2 composite performance and the Trail Making Task ( $r=-.64$ ).

No significant correlations emerged between any UPT-2 variables and the BDEFS-CA. These associations were expected as they are consistent with findings from Wanstall (2019) which used the same sample as the current study. Put together, these findings support that the UPT-2 may be assessing a similar construct as performance-based measures of EF, but not behavioral rating scales, when a community sample of high performing students is assessed. Finally, as hypothesized, there were no significant associations between the performance-based tasks and the behavioral rating scale of EF. This finding supports research highlighting incongruence between EF constructs measured in performance-based tasks and rating scales (Toplak et al., 2013).

### **Associations Among UPT-2 Performance and Academic Abilities**

Measures of academic achievement in math and language domains (raw scores) were significantly associated with performance on the UPT-2. Further, these academic abilities entered as a predictor of UPT-2 performance, before and after controlling for age. Consistent with findings from Wanstall et al. (in preparation), this suggests the contribution of core academic skills in the successful completion of the UPT-2 in a community sample.

### **Associations Among UPT-2 Performance and Age**

Self-directed executive functioning has been defined in the literature as requiring the child to determine their goal, how to reach it, and when (Barker et al., 2014). Associations among UPT-2 performance and age suggest that younger children may be less adept at self-direction, which is consistent with literature on self-directed EFs increasing with age (Ardila et al., 2005, Kavé, 2006, Kavé et al., 2008, Matute et al., 2004, Riva et al., 2000, Sauzéon et al., 2004). Further, these findings replicate those in Ledochowski et al. (2019) and Wanstall (2019). As the UPT-2 develops, future research should work to identify optimal performance in each age or grade group to better understand the development of self-directed EFs and contribute to the development of age or grade-based norms.

### **Metacognitive Ratings of Effort**

The current study found significant correlations within metacognitive effort ratings that met our original hypotheses. Effort exerted was positively associated with ratings of task difficulty and effort required which suggests that the mobilization of effort is determined by the demands of the task (Bambrah et al., 2019). A potential explanation for the non-significant correlations within metacognitive ratings of effort are that evaluation of task difficulty and effort

required concern the objective demands of a task, whereas effort exerted, and the affective experience of effort are subjective interoceptive and emotional states (Naccache et al., 2005).

From a developmental perspective, it is not surprising that ratings of task-based characteristics (task difficulty, effort required) significantly decreased as age increased. Similarly, it is not surprising that ratings of effort and affective experience of effort remained steady as age increases, with the assumption that they reflect an individual characteristic and not a task-dependent characteristic. This conceptualization is limited to the community sample of elementary school students and may look different in a clinical sample of participants.

There were significant gender differences in ratings of effort exerted, effort required, and the affective experience of effort. Research on gender difference in emotional responsivity points to a potential explanation for this finding, such that men have more intense emotional experiences, but females tend to have higher emotional expressivity, particularly for negative emotions (Deng et al., 2016). Future research on the UPT-2 would benefit from an objective measure of effort, for example, a cortisol sample or heart rate monitor assessed while the participant completes the UPT-2. This would allow a better understanding of the experience versus expressivity seen between males and females.

## **UPT-2 Performance and Metacognitive Ratings of Effort**

### *Associations*

All hypotheses were met according to several Spearman's correlations. First, those participants who consider the UPT-2 to be a difficult task, possibly due to its tediousness, performed less well. Second, there was a significant skew of effort exerted ratings, which is likely a factor of the question's formatting. Specifically, the participant is asked to rate how hard they tried. Rather than a high rating indicating they "tried very hard", a high rating represented

“tried my best”. More than 70% of participants rated their effort exerted and 4 or higher on a 5-point Likert scale, which likely suggest that participants rated based on trying their best, not trying their hardest. Third, participants who rated the UPT-2 as requiring higher effort performed less well on the UPT-2 than peers who indicated less effort required. Finally, there were no significant correlations between UPT-2 performance and the affective experience of effort. This is likely due to the high-achieved students within the current study sample.

Spearman correlations between UPT-2 composite performance and metacognitive effort ratings according to gender were completed to better understand the value of these ratings. Similar to the significant differences in ratings, we found that correlations differed according to gender. More specifically, males endorsed trying their best when the task appeared difficult, trying their best when the task required more effort, feeling positive after exerting effort when they tried their best. A marginal and unexpected finding was that males who performed better on the UPT-2 reported that the affective experience effort was negative.

### ***Multiple Regression***

Task difficulty and effort required emerged as significant predictors of UPT-2 performance when entered individually, however, only effort required entered as a significant predictor when all four effort ratings were included in the model. Exploratory findings suggest that when performance on the UPT-2 is divided using a median split, there are significant differences in metacognitive effort ratings depending on UPT-2 performance level (low vs. high performance). This finding may illustrate what results could look like in a clinical vs. control sample (clinical=lower performance level, control=higher performance level). For example, Hsu et al. (2017) reported higher mental effort and discomfort for individuals at-risk for ADHD when compared to individuals not at-risk for ADHD. However, it is important to consider that

performance on the UPT-2 is significantly influenced by the participant's age. In the current study, metacognitive effort ratings differed across grades, and metacognitive effort ratings differed depending on UPT-2 performance level, but metacognitive effort ratings did not differ by performance level when age was held constant. This suggests that there is a strong developmental component to both metacognitive ratings of effort and performance on the UPT-2, thus the current study is unable to isolate these variables from grade/age. Future studies would benefit from isolating performance from age by recruiting a clinical/control sample.

The majority of studies on mental effort have reported a negligible relationship between subjective and objective measures (e.g., Critchley et al., 2004; Ferentzi et al., 2018; Garfinkel et al., 2015). Murphy et al. (2020) suggests the discrepancies as potentially being “driven by the fact that one's self-reported attention may not be predictive of one's objective accuracy”. In the current study, our metacognitive ratings of effort are self-reported attention, whereas objective UPT-2 performance is an objective accuracy and completeness composite variable. Future research would benefit from including rating questions that assess accuracy and attention separately. For example, to measure accuracy we might add ratings of “how many questions do you believe you answered correctly?” and “did you remember to follow instructions?” and compare them to objective analyses of total correct and total complete items on the UPT-2. To measure attention, we might add objective measures of a heart rate monitor and a cortisol sample and compare them to metacognitive ratings of “how did using brainpower make you feel?”.

### **Hierarchical regression predicting UPT-2 performance**

Given the predictive ability of the effort required rating, we chose to examine this rating item alongside other predictors of the UPT-2 including performance-based tasks of EF and a behavioral rating scale of EF. The first hierarchical regression was conducted to determine to

contributions of performance-based measures and the effort required rating in predicting UPT-2 performance. Results indicated that after controlling for age, only performance-based tasks (composite variable) were a significant predictor of UPT-2 composite performance. This finding suggests that the composite variable of performance-based EF tasks is a stronger predictor of UPT-2 composite performance than the effort required rating after controlling for age.

The second regression was conducted to determine whether a behavioral rating scale of EF or the effort required rating was a stronger predictor of UPT-2 performance. Given that age did not correlate with the behavioral rating scale of EF, we chose to omit age in this regression. The multiple regression found that when both ratings (BDEFS-CA and effort required rating item) were entered, the only significant predictor was the effort required rating. More specifically, a single rating item with a specific reference point emerged as a predictor of UPT-2 composite performance, whereas a full behavioral rating scale without specific reference points did not. This finding alludes to the importance of specific reference points in a rating scale when examining performance on an unstructured task.

### **Hierarchical Regressions predicting Academic Abilities**

Two hierarchical regressions were conducted to understand the contributions of performance-based tasks and a behavioral rating scale of EF in predicting academic abilities above and beyond age, UPT-2 composite performance, and the effort required rating. Wanstall et al. (in preparation) indicated that age and UPT-2 composite performance significantly predicted reading abilities, but performance-based tasks of EF and the behavioral rating scale of EF did not. Following the emergence of the effort required rating as a significant predictor of UPT-2 composite performance, we added this variable to the current regression to determine its value in predicting reading abilities when alongside performance-based tasks and a behavioral rating



scale of EF. Thus, the first hierarchical regression was conducted to predict reading abilities. After controlling for age, UPT-2 composite performance, and the effort required rating, only performance-based tasks of EF emerged as a significant predictor of reading abilities. This suggests that performance-based tasks of EF are tapping into a component of a student's reading ability that is not assessed by the behavioral rating scale of EF, above and beyond age, UPT-2 composite performance, and the effort required rating.

Similarly, Wanstall et al. (in preparation) reported that age, UPT-2 composite performance, and the rating scale of EF significantly predicted reading abilities, but performance-based tasks of EF did not. Following the emergence of the effort required rating as a significant predictor of UPT-2 composite performance, we added this variable to the current regression to determine its value in predicting math abilities when alongside performance-based tasks and a rating scale of EF. Thus, an identical hierarchical regression was conducted to predict math abilities. After controlling for age, UPT-2 composite performance, and the effort required rating, neither performance-based tasks of EF nor a rating scale of EF emerged as significant predictors. Notably, age, UPT-2 composite performance, and effort required appeared to be tapping into a component of a student's math ability that is not assessed by performance-based tasks or a rating scale of EF. Finally, the UPT-2 and the effort required rating did not enter as significant predictors of reading abilities, whereas they emerged as significant predictors of math abilities. Although we were not able to replicate Wanstall et al.'s (in preparation) finding that UPT-2 performance is a significant predictor of word reading abilities, we were able to replicate UPT-2 performance as a predictor of math abilities. This finding strengthens Wanstall et al.'s (in preparation) proposal that the UPT-2 may be a valuable tool to help us better understand the development of academic abilities in elementary-aged children.

Bambrah et al. (2019) outline two distinct interpretations of an individual's mobilization of mental effort. The first account suggests that the mobilization of effort is "determined by task demands rather than by the performer's intentions". Their study furthers this conceptualization by suggesting the role of motivation and interest in a performance-based task are significant determinants in mental effort mobilization (e.g., increasing effort as needed, slowing their decline in performance, and regaining optimal performance post-distraction) (Bambrah et al., 2019). This broader account of the mobilization of mental effort may very well play a role in the current study, such that their metacognitive ratings of effort may be influenced by the aforementioned factors. Further, Bambrah et al. (2019) indicate that mental effort is a continuously changing experience. The current study is therefore limited by the ratings only being completed post-task (i.e., recollection ratings) and we might anticipate different effort ratings had they been asked following each item, or similar to Bambrah et al. (2019), prior to, during, and following the UPT-2. Additionally, further research on the UPT-2 might benefit from asking a motivation-related question prior to the task such as "how interested are you in completing this task". Another method to isolate the role of motivation would be to introduce a reward to one group and have a non-reward control group. This would allow for a better understanding of the role that motivation and interest plays in metacognitive ratings of effort.

**Implication of Metacognitive Ratings of Effort** Despite not being a rating scale of EF, the metacognitive ratings of effort were associated with performance-based tasks of EF and performance on the UPT-2, whereas the BDEFS-CA was not. These findings help to illustrate potential significance of having specific reference points when assessing self-reported EF abilities. Just as performance can depend on many factors including structure (instruction, guidance, etc.), ratings of EFs are likely to differ depending on the situation the parents or

teachers are evaluating, for example, during an activity that is tedious compared to a sport they enjoy playing. By evaluating EFs with a rating scale that asks general questions, it omits the possibility for variable performance depending on factors like structure. Specific reference points can guide researchers and practitioners to better understand the role of factors like structure and how they contribute to an individual's abilities. For example, if a version of the UPT-2 were administered with guidelines, item numbers, and equally spaced questions with instructions for each item, and the original UPT-2 were administered, an individual who put forth their best effort on each but finds the unstructured task required more effort may indicate that they benefit from increased structure. Future studies might consider asking children to complete metacognitive ratings of effort following a number of performance-based tasks that range in structure to better understand the value that they place on structure. While the current study was limited to a high performing sample, the metacognitive ratings of effort may contribute more strongly to research that examines clinical populations.

**Implications for Clinical assessment and education** Despite the high performing sample in the current study, clinical implications exist for the UPT-2 and metacognitive ratings of effort. For example, consider the clinical assessment of a student who struggles with their EF abilities (for example, is showing signs of ADHD) but performs well on typical performance-based tasks of EF. Toplak et al. (2013) propose that typical performance-based tasks of EF do not replicate the EFs that are used in daily living, which may explain why the students fared well on a highly structured task. Further research on the UPT-2 and metacognitive ratings of effort with clinical and control samples are required. If research were to suggest that performance on structured and unstructured EF tasks varies in clinical populations but not in control samples, clinicians could benefit from administering the UPT-2 in addition to a more structured

performance-based task and inquire about metacognitive ratings of effort for each task. In addition to its potential for adjunct diagnostic utility, stronger performance on measures with more structure might indicate that the integration of structure may be a helpful accommodation for the student to reach their full potential. Conceptually, structure might be offered variably in a class as needed.

The current study has implications for learning regulation in classrooms. Faith (2022) discusses the “Barriers & Strategies Protocol” (BSP) as a classroom tool to share “task understanding (“What are our barriers to this task?”), devise cognitive strategies (“What strategies could we use to be successful?”), and report their responses in a t-chart”. The BSP was developed with the intention of shifting away from typical teacher-directed behaviors and promoting student metacognition by collecting context specific information about student challenges (Faith, 2022). Similar to the BSP’s context specific information, the metacognitive ratings in the current study are comprised of context specific items which refer to the UPT-2. The UPT-2 and metacognitive ratings of effort invite dialogue surrounding barriers and cognitive strategies in assessment measures, daily activities, and classroom tasks. If students are aware and understand what structure might entail (instruction, time limits, practice questions, feedback, etc.) they are provided with a resource to reflect on when a task is difficult and communicate whether a different degree of structure may better suit their needs. For example, a “barrier” to the UPT-2 may be its scattered questions which require self-direction and goal-orientation, and the “cognitive strategies” may entail completing questions left to right, by domain (math, language, symbolic), by perceived difficulty. Given the relatively easy nature of the UPT-2, a child’s rating of task difficulty (or other ratings) may be helpful to understand in terms of their understanding of task structure and whether they believe additional structure would improve their performance.

### **Limitations in the Current Study**

There are important limitations to consider in this project. First, the community sample used in this study was considered to be a high-performing sample. The participating school holds a philosophy which values inquiry-based learning and a focus on self-monitoring of skills and abilities. Future research would benefit from replicating this work on the UPT-2 in more diverse settings such as public schools, psychology clinics, and hospitals. Secondly, only the Stroop and Trail Making tasks were administered to assess EF performance. Future studies may include additional performance-based measures of EF to examine their relationship with the UPT-2. Third, rating-based measures are subject to a degree of interpretation. In terms of the BDEFS, the lack of specific reference point leaves the parent to decide the context they consider when answering an item. In terms of the metacognitive ratings of effort, children are provided with a specific reference point to consider when answering an item, however, implicit theories of effort and intelligence may play a role in how a student chooses to answer rating items. Two implicit theories of intelligence have been suggested in the literature: (1) a growth mindset believes that intellectual ability is malleable, and (2) a fixed mindset (Dweck & Leggett, 1988). These theories have implications for cognitive effort, such that individuals with a growth mindset maintain that the amount of effort put into a task is reflected in your performance, whereas those with a fixed mindset do not believe that effort is related to task performance (Scheiter et al., 2020). Put together, interpretations of effort may vary depending on their conceptualization of intelligence. Although the current study did not include identifications of alignment with a growth or fixed mindset, future research on the UPT-2 and metacognitive ratings of effort would benefit from including a measurement of participants' theory of intelligence to better understand the predictive ability of effort ratings in student's with a growth vs. fixed mindset. Finally, when we

hypothesize that effort ratings are predictors of UPT-2 performance, we are under the assumption that these participants have the same level of metacognition as their peers. A number of factors have been found to influence an individual's metacognitive processes, namely cognitive ability, affective function, physical health, and mental health (Brewer et al., 2015; Craig, 2003; Garfinkel et al., 2015; Khalsa et al., 2018; Murphy et al., 2018; Quattrocki & Friston, 2014). Future research would benefit from interpreting performance on the UPT-2 and metacognitive ratings of effort in the context of these individual differences.

### **Conclusion**

The current study examined a novel Unstructured Performance-Based Task (UPT-2) in a community sample of children. Significant associations were found between the UPT-2 and performance-based tasks, rating scale, and academic abilities. Additionally, there were no significant associations between the rating scale and either performance-based task. In addition to significant associations between the UPT-2 and metacognitive ratings of effort, the ratings significantly predicted UPT-2 performance. Significant associations within metacognitive ratings of effort were found. Further analyses determined that performance-based tasks were better predictors of UPT-2 composite performance than a metacognitive rating of effort, and a metacognitive rating of effort was a better predictor of UPT-2 composite performance than a rating scale of EF. Finally, age, the UPT-2, and the effort rating emerged as stronger predictors of reading and math abilities over the rating scale of EF, and of math abilities over the performance-based tasks of EF. Overall, these results indicate the UPT-2 may be a promising measure to assess EF related difficulties and the addition of metacognitive ratings of effort act as a complementary component to better understanding a student's performance and how to best support them.

## References

- Ackerman, R., & Thompson, V. A. (2017). Meta-Reasoning: monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, 21(8), 607–617.  
<https://doi.org/10.1016/j.tics.2017.05.004>.
- Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. *Child Neuropsychology*, 8(2), 71–82. <https://doi.org/10.1076/chin.8.2.71.8724>
- Anderson, P. (2008). Towards a developmental model of executive function. In V. Anderson, R. Jacobs & P. J. Anderson (Eds.), *Executive functions and the frontal lobes: A lifespan perspective; executive functions and the frontal lobes: A lifespan perspective* (pp. 3-21, Chapter xxxiii, 541 Pages) Taylor & Francis, Philadelphia, PA. Retrieved from <http://ezproxy.library.yorku.ca/login?url=https://www-proquest-com.ezproxy.library.yorku.ca/books/towards-developmental-model-executive-function/docview/621807033/se-2?accountid=15182>
- Ardila, A., Rosselli, M., Matute, E., & Guajardo, S. (2005). The influence of the parents' educational level on the development of executive functions. *Developmental Neuropsychology*, 28, 539–560.
- Bambrah, V., Hsu, C.-F., Toplak, M. E., & Eastwood, J. D. (2019). Anticipated, experienced, and remembered subjective effort and discomfort on sustained attention versus working

memory tasks. *Consciousness and Cognition*, 75, 102812.

<https://doi.org/https://doi.org/10.1016/j.concog.2019.102812>

Barkley, R. A. (2012). *Barkley deficits in executive functioning scale—children and adolescents (BDEFS-CA)*. New York, NY: Guilford Press.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, 64, 417–444.

<https://doi.org/10.1146/annurev-psych-113011-143823>

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

Brewer, R., Happé, F., Cook, R., & Bird, G. (2015). Commentary on “Autism, oxytocin and interoception”: Alexithymia, not Autism Spectrum Disorders, is the consequence of interoceptive failure. *Neuroscience and Biobehavioral Reviews*, 56, 348–353.

doi:10.1016/j.neubiorev.2015.07.006

Brown, T. E. (1996). *Brown Attention-Deficit Disorder Scales: Adolescents and Adults*.

Craig, A. D. (2003). Interoception: the sense of the physiological condition of the body. *Current Opinion in Neurobiology*, 13(4), 500–505. doi:10.1016/S0959-4388(03)00090-4

Critchley, H. D., Wiens, S., Rotshtein, P., Öhman, A., & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, 7, 189–195.

Deng, Y., Chang, L., Yang, M., Huo, M., & Zhou, R. (2016). Gender differences in emotional response: Inconsistency between experience and expressivity. *PLoS ONE*, 11(6), 1–12.

<https://doi.org/10.1371/journal.pone.0158666>



Diamond, A. (2013). Executive functions. *Annual review of psychology*, 64, 135–168.

<https://doi.org/10.1146/annurev-psych-113011-143750>

Diamond, A. (2020). Chapter 19 - Executive functions. In A. Gallagher, C. Bulteau, D. Cohen, & J. L. Michaud (Eds.), *Neurocognitive Development: Normative Development* (Vol. 173, pp. 225–240). Elsevier. <https://doi.org/https://doi.org/10.1016/B978-0-444-64150-2.00020-4>

Diamond, A., Ling, D.S., (2016). Conclusions about interventions, programs, and approaches for improving executive functions that appear justified and those that, despite much hype, do not. *Developmental Cognitive Neuroscience*. 18, 34–48.

<http://dx.doi.org/10.1016/j.dcn.2015.11.005>

Doty, S. J., Hixson, M. D., Decker, D. M., Reynolds, J. L., & Drevon, D. D. (2015). Reliability and Validity of Advanced Phonics Measures. *Journal of Psychoeducational Assessment*, 33(6), 503–521. <https://doi.org/10.1177/0734282914567870>

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, 95,256–273. <https://doi.org/10.1037/0033-295x.95.2.256>

Faith, L. C. (2022). *Achieving Learning Regulation Support: How a Socially Shared Approach Fits, Functions, and Thrives* (Order No. 28773328). Available from ProQuest Dissertations & Theses Global. (2645185806).

<https://ezproxy.library.yorku.ca/login?url=https://www.proquest.com/dissertations-theses/achieving-learning-regulation-support-how/docview/2645185806/se-2?accountid=15182>

- Ferentzi, E., Drew, R., Tihanyi, B. T., & Köteles, F. (2018). Interoceptive accuracy and body awareness—Temporal and longitudinal associations in a non-clinical sample. *Physiology & Behavior, 184*, 100–107. doi:10.1016/j.phys-beh.2017.11.015
- Garfinkel, S. N., Seth, A. K., Barrett, A. B., Suzuki, K., & Critchley, H. D. (2015). Knowing your own heart: distinguishing interoceptive accuracy from interoceptive awareness. *Biological Psychology, 104*, 65–74. doi:10.1016/j.biopsycho.2014.11.004
- Gioia, G. A., Isquith, P. K., & Kenealy, L. E. (2008). Executive functions and the frontal lobes: A lifespan perspective. New York: Taylor & Francis.
- Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2015). BRIEF-2: Behavior rating inventory of executive function. Lutz, FL: Psychological Assessment Resources.
- Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). BRIEF – Behavior rating inventory of executive function. Professional manual. Odessa, FL: Psychological Assessment Resources.
- Golden, C. J. (1978). Stroop colour and word test. *Age, 15*, 90.
- Heaton, R.K., Chelune, G.J., Talley, J.L., Kay, G.G., & Curtis, G. (1993). Wisconsin Card Sorting Test (WCST) manual, revised and expanded. Odessa, FL: Psychological Assessment Resources.
- Hsu, C. F., Eastwood, J. D., & Toplak, M. E. (2017). Differences in perceived mental <https://doi.org/10.3389/fpsyg.2017.00407>
- Jensen, A.R., & Rohwer, W.D. Jr. (1966). The Stroop color- word test: A review. *Acta Psychologica, 25*, 36–93
- Kavé, G. (2006). The development of naming and word fluency: Evidence from Hebrew-speaking children between ages 8 and 17. *Developmental Neuropsychology, 29*, 493–508.

- Kavé, G., Kigel, S., & Kochva, R. (2008). Switching and clustering in verbal fluency tasks throughout childhood. *Journal of Clinical and Experimental Neuropsychology*, *30*, 349–359.
- Khalsa, S. S., Adolphs, R., Cameron, O. G., Critchley, H. D., Davenport, P. W., Feinstein, J. S., ... Interoception Summit 2016 participants. (2018). Interoception and mental health: A roadmap. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(6), 501–513. doi:10.1016/j.bpsc.2017.12.004
- 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 and Experimental Neuropsychology*, *41*(5), 445–459.  
<https://doi.org/10.1080/13803395.2019.1567694>
- Li, W., Lee, A., & Solmon, M. (2007). The role of perceptions of task difficulty in relation to self-perceptions of ability, intrinsic value, attainment value, and performance. *European Physical Education Review*, *13*(3), 301–318. <https://doi.org/10.1177/1356336X07081797>
- MacLeod, C.M. (1991). Half a century of research on the Stroop Effect: An integrative review. *Psychological Bulletin*, *109*, 163–203
- Mahone, E. M., Cirino, P. T., Cutting, L. E., Cerrone, P. M., Hagelthorn, K. M., Hiemenz, J. R., ... Denckla, M. B. (2002). Validity of the behavior rating inventory of executive function in children with ADHD and/or Tourette syndrome. *Archives of Clinical Neuropsychology*, *17*(7), 643–662.
- Matute, E., Rosselli, M., Ardila, A., & Morales, G. (2004). Verbal and nonverbal fluency in Spanish-speaking children. *Developmental Neuropsychology*, *26*, 647–660.

- Maynard, D. C., & Hakel, M. D. (1997). Effects of Objective and Subjective Task Complexity on Performance. *Human Performance*, *10*(4), 303–330.  
[https://doi.org/10.1207/s15327043hup1004\\_1](https://doi.org/10.1207/s15327043hup1004_1)
- 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 Neuropsychological Society*, *16*(3), 495–505. [PubMed: 20188014]
- 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. <https://doi.org/https://doi.org/10.1006/cogp.1999.0734>
- Murphy, J., Brewer, R., Plans, D., Khalsa, S. S., Catmur, C., & Bird, G. (2020). Testing the independence of self-reported interoceptive accuracy and attention. *Quarterly Journal of Experimental Psychology*, *73*(1), 115–133. <https://doi.org/10.1177/1747021819879826>
- Murphy, J., Catmur, C., & Bird, G. (2018). Alexithymia is associated with a multidomain, multidimensional failure of interoception: Evidence from novel tests. *Journal of Experimental Psychology. General*, *147*(3), 398–408. doi:10.1037/xge0000366
- Naccache, L., Dehaene, S., Cohen, L., Habert, M.-O., Guichart-Gomez, E., Galanaud, D., & Willer, J.-C. (2005). Effortless control: executive attention and conscious feeling of mental effort are dissociable. *Neuropsychologia*, *43*(9), 1318–1328.  
<https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2004.11.024>
- O’Brien, A. M., Kivisto, L. R., Deasley, S., & Casey, J. E. (2021). Executive Functioning Rating Scale as a Screening Tool for ADHD: Independent Validation of the BDEFS-CA.

- Journal of Attention Disorders*, 25(7), 965–977.  
<https://doi.org/10.1177/1087054719869834>
- Otto, T., Zijlstra, F., & Goebel, R. (2014). Neural correlates of mental effort evaluation— involvement of structures related to self-awareness, *Social Cognitive and Affective Neuroscience*, 9(3), 307-315. <https://doi.org/10.1093/scan/nss136>
- Peng, Y., & Tullis, J. G. (2020). Theories of intelligence influence self-regulated study choices and learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(3), 487–496. <https://doi.org/10.1037/xlm0000740>
- Pennington, B. F., & Ozonoff, S. (1996). Executive functions and developmental psychopathology. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 37(1), 51–87. <https://doi.org/10.1111/j.1469-7610.1996.tb01380.x>
- Petscher, Y., & Logan, J. (2014). Quantile regression in the study of developmental
- Quattrocki, E., & Friston, K. (2014). Autism, oxytocin and interoception. *Neuroscience and Biobehavioral Reviews*, 47, 410–430. doi:10.1016/j.neubiorev.2014.09.012
- Reitan, R. M. (1955). The relation of the Trail Making Test to organic brain damage. *Journal of Consulting Psychology*, 19, 393-394. doi:10.1037/h0044509
- Reitan, R. M. (1958). Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and Motor Skills*, 8, 271-276. doi:10.2466/PMS.8.7.271-276  
 San Antonio, TX: Psychological Corporation.
- Riva, D., Nichelli, F., & Devoti, M. (2000). Developmental aspects of verbal fluency and confrontation naming in children. *Brain and Language*, 71, 267–284.
- Sauz on, H., Lestage, P., Raboutet, C., N’Kaoua, B., & Claverie, B. (2004). Verbal fluency output in children aged 7–16 as a function of the production criterion: Qualitative

- analysis of clustering, switching processes, and semantic network exploitation. *Brain and Language*, 89, 192–202.
- Scheiter, K., Ackerman, R., & Hoogerheide, V. (2020). Looking at Mental Effort Appraisals through a Metacognitive Lens: Are they Biased? *Educational Psychology Review*, 32(4), 1003–1027. <https://doi.org/10.1007/s10648-020-09555-9>
- Schrank, F. A., Mather, N., McGrew, K. S. (2014). Woodcock-Johnson IV Tests of Achievement. Rolling Meadows, IL: Riverside.
- sciences. *Child development*, 85(3), 861–881. <https://doi.org/10.1111/cdev.12190>
- Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological Science*, 29(4), 549–571. <https://doi.org/10.1177/0956797617739704>
- Snyder, H. R., & Munakata, Y. (2013). So many options, so little control: Abstract representations can reduce selection demands to increase children’s self-directed flexibility. *Journal of Experimental Child Psychology*, 116(3), 659–673. <https://doi.org/10.1016/j.jecp.2013.07.010>
- Stanovich, K. E. (2009). “Distinguishing the reflective, algorithmic, and autonomous minds: is it time for a tri-process theory?,” in *In Two Minds: Dual Process and Beyond*, eds J. Evans and K. Frankish (Oxford: Oxford University Press), 55–88.
- Steyer, R., Ferring, D. & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*, 8, 79–98.

- Steyer, R., Majcen, A.-M., Schwenkmezger, P. & Buchner, A. (1989). A latent state-trait anxiety model and its application to determine consistency and specificity coefficients. *Anxiety Research, 1*, 281–299.
- Strauss, E., Sherman, E.M.X., & Spreen, O. (2006). A compendium of neuropsychological tests (3rd edn). New York: Oxford University Press
- Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–662
- 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 and Allied Disciplines, 54*(2), 131–143.  
<https://doi.org/10.1111/jcpp.12001>
- Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2012). Test of Word Reading Efficiency-Second Edition. Austin, TX: Pro-Ed
- Villarreal, V. (2015). Test Review: Schrank, F. A., Mather, N., & McGrew, K. S. (2014).
- Wanstall, E. (2019) *Unstructured performance task to assess executive functions: a study in typically-developing children*. [Master's thesis, York University]. Yorkspace.
- Wanstall, E., Martinussen, R., Toplak, M. E. (in preparation). *Predicting academic performance from a novel task of self-direction in elementary-aged students*.
- Woodcock-Johnson IV Tests of Achievement. *Journal of Psychoeducational Assessment, 33*(4), 391–398. <https://doi.org/10.1177/0734282915569447>
- Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., & Weintraub, S. (2013). II. NIH Toolbox Cognition Battery (CB): measuring executive function and

attention. *Monographs of the Society for Research in Child Development*, 78(4), 16–33.


















<https://doi.org/10.1111/mono.12032>



## List of Appendices

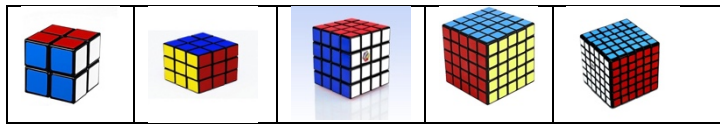
Appendix A: Unstructured Performance Task – 2 <sup>nd</sup> Version.....	62
Appendix B: Metacognitive Ratings of Effort.....	63

## Appendix A. Unstructured Performance Task-2<sup>nd</sup> Version

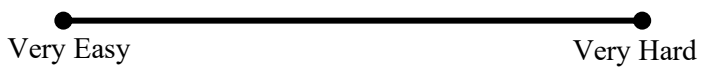
<p></p> <p>2, 6, 5, 9, 8, ___</p>	<p style="text-align: center;">Write the first letter of your name.</p> <p style="text-align: center;">1 2 3 4 5 ___</p>	<p style="text-align: right;">1 2 3 4 5 ___</p>
<p></p> <p>4+3 =</p>	<p style="text-align: center;">Do "pot" and "not" rhyme?</p> <p style="text-align: center;">8 + 4 + 3 =</p> <p style="text-align: center;">B D G B D G B ___</p> <p style="text-align: center;">Do through and though rhyme?</p>	<p style="text-align: center;">C G H H C G H ___</p> <p style="text-align: right;">Is this a dog? </p> <p style="text-align: center;">Do bark and part rhyme?</p> <p style="text-align: right;">13 - 5 =</p>
<p>Write a word that ends with the letter M.</p> <p>↑ → ↑ → ↑</p>	<p> _ i s h</p> <p style="text-align: center;">Do airplanes have wings?</p>	<p></p> <p style="text-align: right;">Draw a rectangle.</p>
<p>Is this a circle? </p> <p>K J F K K J F K ___</p>	<p style="text-align: center;">Circle the picture that does not fit.</p> <p></p> <p>15, 12, 9, 6, ___</p>	<p style="text-align: right;"><b>TURN OVER TO CONTINUE!</b></p> <p style="text-align: right;"></p>
<p>What is this? </p> <p>Is this a bug? </p>	<p>24 - 11 - 2 =</p> <p>8 - 6 =</p> <p>Birds live in a ___.</p>	<p style="text-align: center;">← ↑ → ← ↑ →</p> <p style="text-align: center;">Write a word that contains the letter E, S, and B.</p> <p style="text-align: center;">Do feet and meat rhyme?</p>
<p>A I I O A I I O ___</p> <p>How many letter t's are in this sentence: This turtle ate tulips.</p>	<p> _ u s</p>	<p></p> <p style="text-align: center;">Circle the picture that does not fit.</p> <p></p>
<p>Colour all of the triangles. </p> <p>Draw a house.</p>	<p>2, 4, 6, 8, ___</p> <p></p>	<p style="text-align: center;">A E A E A E A ___</p> <p style="text-align: right;">12 + 17 =</p> <p style="text-align: right;">3, ___, 13, 18</p>
<p>Draw a dot in each circle. </p> <p>1+2 =</p>	<p style="text-align: center;">How many "sh" sounds are in this sentence: She wished for a shiny new bicycle.</p> <p style="text-align: center;">6 5 4 3 ___</p>	<p style="text-align: right;"><b>TURN OVER TO CONTINUE!</b></p> <p style="text-align: right;"></p>
<p>A word that rhymes with "bat" ___</p>		

Appendix B. Metacognitive Ratings of Effort

1) How **hard** was this activity?



1                      2                      3                      4                      5



2) How much did you **try** on this activity?

1                      2                      3                      4                      5



3) How much of your **brainpower** did this activity require?



1    2    3



4) How did using **brainpower** make you **feel**?

Point to the face that best shows how you feel after doing this activity.

