Journal of Experimental Child Psychology 222 (2022) 105462

Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

journal homepage: www.elsevier.com/locate/jecp

Converging evidence for domain-general developmental trends of mental attentional capacity: Validity and reliability of full and abbreviated measures

Adriana Milani [a], Juan Pascual-Leone [a], Marie Arsalidou [a,b,*]

[a] Department of Psychology, York University, North York, Ontario M3J 1P3, Canada
[b] Department of Psychology, National Research University Higher School of Economics (HSE University), Moscow 101000, Russian Federation

ARTICLE INFO

Article history:
Received 5 July 2021
Revised 20 April 2022

Keywords:
Mental attentional capacity
Cognitive development
Contextual influences
Abbreviated measures
Parametric tasks

ABSTRACT

Our ability to understand the world around us hinges on our cognition. Theoretically, children's abilities improve with age; however, a lively discussion exists on how factors such as task domain, task interference, and task difficulty, as indexed mainly by relevant cues, affect cognitive performance. Practically, cognitive measures take a substantial amount of time to administer, which poses limitations for researchers in the field of psychology. The current study addressed theoretical and practical questions regarding the nature of child cognition using full and abbreviated versions of classic, recent, and new tasks of mental attentional capacity. We employed a cross-sectional design testing 483 participants in six groups (7–30 years of age) on the new Number Matching Task (NMT), the established Color Matching Task (CMT), and the classic Figural Intersection Task (FIT). Results confirm theoretical predictions of the developmental increase in mental attentional capacity and the adjunct hypothesis that tasks with high interference are better to assess the developmental trajectory of mental attentional capacity quantitatively. NMT scores are significantly equivalent to CMT scores and the theoretically predicted mental attentional capacity, and the abbreviated CMT and NMT produce comparable scores to those of the full tests. We determined that the NMT can be administered developmentally and is

---

* Corresponding author at: Department of Psychology, York University, North York, Ontario M3J 1P3, Canada.
E-mail addresses: arsalido@yorku.ca, marsalidou@hse.ru (M. Arsalidou).

https://doi.org/10.1016/j.jecp.2022.105462

appropriate for use in assessing mental attentional capacity in studies with both children and adults.

Introduction

Children's cognitive abilities improve with age, and the mechanisms that bring about this change have been an influential catalyst for research in experimental child psychology. Many developmental theories provide qualitative descriptions of cognitive changes across development (Case, 1992; Demetriou & Spanoudis, 2018; Halford, Cowan, & Andrews, 2007; Halford, Wilson, & Phillips, 1998; Pascual-Leone, 1970); however, few quantify them (e.g., Halford et al., 2007; Pascual-Leone, 1970; Pascual-Leone & Johnson, 2021). The theory of constructive operators proposes mechanisms of developmental changes as they relate to mental attention (e.g., Pascual-Leone, 1970; Pascual-Leone & Johnson, 2021). Mental attentional capacity is a core domain-general cognitive resource that corresponds to the maximum energy an individual can mobilize during problem solving.

Seven-year-olds are expected to have on average three units of mental attentional capacity, which increases by one unit every other year, and by the time they are 16 years old this number has increased to seven units. Many tasks of mental attentional capacity are designed to assess the developmental trajectory of mental attentional capacity (e.g., Figural Intersection Task: Pascual-Leone & Baillargeon, 1994; Direction Following Task: Pascual-Leone & Johnson, 2011). Tasks of mental attention have various characteristics, including multiple levels of difficulty (i.e., features/cues that need to be considered during problem solving), invariant executive rules (i.e., the rule of the task remains unchanged regardless of difficulty level), and controlled interference. The Color Matching Tasks are visual–spatial tasks of mental attention are computerized and have been used with functional neuroimaging (Arsalidou, Pascual-Leone, & Johnson, 2010; Arsalidou, Pascual-Leone, Johnson, Morris, & Taylor, 2013; Vogan et al., 2014; Vogan, Morgan, Powell, Smith, & Taylor, 2016). We are not aware of a task that assesses mental attentional capacity in the numeric domain. In this study, we developed the Number Matching Tasks, new numeric measures that assess mental attentional capacity that can be used behaviorally and with functional neuroimaging.

The purpose of the current study had three main goals: (a) investigate task properties in measures of mental attentional capacity (i.e., task domain, difficulty, and interference) used to assess cognitive abilities across development, (b) replicate past findings in established measures and validate a new measure of mental attentional capacity, and (c) evaluate the number of trials needed to obtain a valid and reliable mental attentional capacity score using these *metasubjective* (i.e., thinking about the task process from a participant's perspective) *measures.* Practically, abbreviated measures would be an asset for future research because they can lead to increased use due to time saving in research, educational, and clinical settings. Furthermore, data from multiple age groups in two domains can be valuable for future quantitative meta-analyses. Theoretically, findings will contribute to better understanding of mechanisms that underlie cognitive development.

We framed our study on the theory of constructive operators (TCO). The TCO was conceived during the 1960 s while Juan Pascual-Leone was studying with Jean Piaget. The first publication of this theory was proposed as a mathematical model for predicting Piaget's developmental stages (Pascual-Leone, 1970). The most recent rendering of this theory appears in a book by Pascual-Leone and Johnson (2021) that covers behavioral and neurobiological foundations of the development of psychological processes. The TCO is a general theory and is influenced by both Piaget's developmental constructivist theory and the theories of Vygotsky and Luria (Pascual-Leone, 1987, 1995, 2012, 2014) as well as neuroscience. Central to the TCO is the developmental construct of mental attentional capacity. Mental attention we believe, is indexed by the number of items, not facilitated by the situation, that one can maintain and manipulate. The TCO and the model of endogenous mental attention propose the actions of operators, schemes, and principles that interact to bring about performance.

Operators are content-free (i.e., general-purpose) regulations that can apply on schemes in any domain (Pascual-Leone, 1970; Pascual-Leone & Johnson, 1991, 2005, 2011). Psychologically, schemes represent information-bearing units that are self-propelled and are classified into three groups: figurative (e.g., distal objects), operative (e.g., actions), and executive (e.g., plans) schemes. The principle of schematic over-determination of performance (SOP; Pascual-Leone & Johnson, 1991, 2005, 2011, 2021) determines which schemes will eventually apply to generate an outcome. It stipulates that any actual performance is the product of *all compatible mobilized schemes available to the situation that converge to the most activated scheme.* This principle derives in part from Sherrington's (1940) idea of a generalized principle of convergence because he wrote, "Where it is a question of 'mind' the nervous system does not integrate itself by centralization upon one pontifical cell. Rather it elaborates a million-fold democracy whose each unit is a cell" (p. 277).

Mental attentional capacity (*M*−capacity) is generally defined as a set of mental processes responsible for active maintenance and manipulation in mind of task-relevant problem-solving information that is not otherwise (e.g., perceptual cueing) activated (e.g., Pascual-Leone, 1970; Pascual-Leone & Johnson, 2005, 2011, 2021). *Mental attention* (*M*) is often described as an attentional resource that serves as a core component of working memory (Arsalidou et al., 2010) because it provides operational mental energy for this processing to occur. Measurement of *M*−capacity allows researchers to determine the type and quantity of otherwise not activated *schemes* (informational units) that can be maintained by mental attention and thus can be manipulated (Arsalidou & Im-Bolter, 2017; Pascual-Leone & Johnson 1999, 2011). In line with predictions of the TCO (Pascual-Leone 1970; Pascual-Leone & Johnson 2005), this capacity has a developmental growth, when assessed behaviorally, of one unit every other year after 3 years of age—until it reaches its maximum of seven units at 15 or 16 years of age (see Arsalidou & Im-Bolter, 2017, for review; Pascual-Leone & Baillargeon, 1994).

Measures of *M*−capacity are often used in developmental, educational, clinical, and cognitive neuroscientific research domains, typically in an effort to predict performance on a wide range of cognitive tasks and real-world outcomes. *M*−capacity is fundamentally involved in language and reading comprehension (Im-Bolter, Johnson, & Pascual-Leone, 2006) and learning and scholastic achievement (Onwumere & Reid, 2008), and it has also predicted individual differences such as performance on standardized intelligence tests (Howard, Johnson, & Pascual-Leone, 2013), mathematical cognition (Agostino, Johnson, & Pascual-Leone, 2010), cognitive giftedness (Johnson, Im-Bolter, & Pascual-Leone, 2003), and theory of mind (Im-Bolter, Agostino, & Owens-Jaffray, 2016). Tasks of *M*−capacity have been used with individuals with clinical diagnoses such as autism (Vogan et al., 2014; Vogan, Morgan, Smith, & Taylor, 2019), and specific language impairments (Im-Bolter et al., 2006). Considering the range and frequency of use in the field, it is of importance to accurately and reliably measure individual and developmental differences in *M*−capacity.

Arsalidou et al. (2010) designed the Color Matching Task (CMT), which measure *M*−capacity developmentally and are appropriate for use in functional neuroimaging studies with both children and adults (e.g., Arsalidou et al., 2013; Vogan et al., 2014, 2019). In the current study, we used the CMT as a model to design the Number Matching Task (NMT). The CMT and NMT are computerized timed paradigms, variants of a match-to-sample task, where difficulty of cognitive processes increases with the number of relevant cues in the stimulus. Participants see a sequence of stimuli, and for each stimulus they must indicate whether the relevant cues (i.e., colors or numbers) are the same as or different from those immediately preceding. The number of relevant colors or numbers varies from one to six and index the level of difficulty for each item. Irrelevant cues (irrelevant colors or numbers) are kept constant across the six levels of difficulty, and participants are instructed to ignore them. All cues are embedded in a figure, which in itself is irrelevant for the task at hand. Common characteristics of the CMT and NMT include (a) multiple levels of difficulty that are parametrically scaled, (b) invariant executive demand across levels of difficulty (i.e., the number and kind of operations needed to solve the task item are kept constant), (c) minimal prior knowledge requirements due to pretraining, (d) minimal language or conceptual requirements, and (e) use of irrelevant cues that evoke interference—either high or low interference. Each task version consists of items with changing low- and high-level interference, manipulated by the number and kind of irrelevant cues present, to produce two different degrees of interference (see Method for more details on task variations). The multifaceted structure of the tasks allows for manipulations of domain (e.g., visual–spatial vs. numeric), difficulty (e.g., multiple levels of difficulty), and interference (high interference vs. low interference). Thus, effect of difficulty will support the developmental trajectory for mental attention, lack of effect of domain will support the notion of a common core cognitive resource used in problem solving, and effects of interference will document the context that elicits the need for additional executive function skills.

One way in which higher-interference variants produce increased contextual interference is by way of relevant cues being embedded in salient irrelevant figures, which distract from the task at hand, making the search for numbers or colors more attentionally demanding because the participants need to disembed the relevant features from the global/embedding figures. In addition, in the numeric higher-interference variant, a large number (global level) is composed of smaller numbers (local level), which produces a Navon effect (Gerlach & Poirel, 2018; Navon, 1977). These factors hinder performance, causing participants to produce more errors, which lowers the probability of success unless they use mental attention (Arsalidou et al., 2010; Pascual-Leone, 1970).

Behavioral research shows that these tasks with embedded contexts and irrelevant cues, which evoke interference, are better able to assess *M*−capacity (Arsalidou et al., 2010; Pascual-Leone & Baillargeon, 1994; Pascual-Leone & Johnson, 2005; Pascual-Leone, Johnson, Baskind, Dworsky, & Severtson, 2000; Powell, Arsalidou, Vogan, & Taylor, 2014). This is one advantage that metasubjective *M*−measures such as the CMT and NMT can offer to the fields of psychology and education. In addition, these *M*−measures (including the CMT and NMT), with their parametrically scaled levels of difficulty, provide researchers with tools for assessing brain–behavior relations in the visual–spatial and numeric domains (see Arsalidou & Im-Bolter, 2017, for review). Although the tasks can be classified by their stimulus type to belong to specific domains (i.e., visual–spatial and numeric), we believe that what is measured by the tasks is a *domain-general resource* for cognition (Pascual-Leone 1970, 1987; Pascual-Leone & Johnson, 2005, 2011). Tasks assessing a common resource in different domains can be advantageous in certain circumstances. Specifically, Im-Bolter et al. (2006) showed that children with specific language impairments have *M*−capacity deficits in both the visual–spatial and verbal domains. Thus, both tests, although measuring a domain-general resource, can be used to assess characteristic (domain-specific) impairments in patients as well as in children with atypical development. These parametric measures have a multitude of critical advantages when used in developmental cognitive neuroscience or neuropsychology and related fields (see Arsalidou & Im-Bolter, 2017, for review).

The CMT and NMT have been validated and reliably used in behavioral studies (Arsalidou et al., 2010; Powell et al., 2014; Vogan et al., 2016) and functional magnetic resonance imaging (fMRI) studies with both healthy samples (Arsalidou et al., 2013; Guevara, 2018) and clinical samples (e.g., Vogan et al., 2014, 2019). Critically, these tasks take a considerable amount of time to administer (i.e., ~40 min each including pretraining). Given that these tests are critical tools of *M*−capacity assessment, researchers and clinicians working in several subfields of psychology would find it very practical if these tests were administered in a shorter time.

Many cognitive and neuropsychological measures have benefited from a substantial shortening of administration time. For instance, complex span tasks have been shortened to meet the time concerns of researchers who use them (Felez-Nobrega, Foster, Puig-Ribera, Draheim, & Hillman, 2018; Foster et al., 2015). Similar criticisms have motivated researchers and clinicians assessing other cognitive functions. In the domain of clinical neuropsychology, where brief cognitive screening measures are so important, shortening tools of assessment has gained momentum. For example, to name just three medical conditions, neurocognitive batteries have been abbreviated to screen for dementia (Horton et al., 2015), schizophrenia (Harvey, Keefe, Patterson, Heaton, & Bowie, 2009), and mild cognitive impairment (Gomar, Harvey, Bobes-Bascaran, Davies, & Goldberg, 2011). Abbreviated tests have become ubiquitous in the field.

The primary motivation to shorten test duration is saving time. The increasing pressures on clinicians to conduct efficient, cost-effective, and short assessments, given reimbursement policies of third-party companies, is long-standing in the field (Camara, Nathan, & Puente, 2000; Wright et al., 2017). In addition, in the research setting long experiments are extremely costly, particularly using neuroimaging tools. Thus, researchers have abbreviated the Penn Verbal Reasoning Test, a computerized measure of verbal reasoning capacity used in neuroimaging investigations (Bilker, Wierzbicki, Brensinger, Gur, & Gur, 2014). Another important reason for shortening tasks of psychological assessment is the participants' motivation or fatigue. Participants' motivation and attentional resources can decline with lengthy and tiresome experimental sessions, influencing performance assessment (Felez-Nobrega et al., 2018; Vanderploeg, LaLone, Greblo, & Schinka, 1997). Abbreviated measures used for evaluation in clinical neuropsychology benefits patients with cognitive impairment because of difficulties associated with fatigue and attentional decline (Gomar et al., 2011). It is imperative for researchers and clinicians to have access to measures that are time-sensitive yet remain psychometrically sound. In this study, we present information on abbreviated measures of mental attentional capacity.

There is no consistent method or practice for developing reliable and valid short forms parallel to an original version. A survey of the literature shows that when abbreviating the same test, different methods can be adopted. For instance, Calamia, Markon, Denburg, and Tranel (2011) used techniques from item response theory to shorten the Judgement of Line Orientation Test, whereas other researchers used odd–even splits (Vanderploeg et al., 1997; Woodard et al., 1996). In addition, in the case of intelligence testing where strategies often consist of reducing the number of subtests or number of items from subtests, different methods have been adopted (see Kleka & Paluchowski, 2017, for review). Specifically, examining the methods used to shorten the scale of questionnaires showed that statistical techniques investigated were comparable (Kleka & Paluchowski, 2017). Hence, they advocated equal treatment of statistical techniques, with the basis for their selection guided by the availability of analytic tools and the preferences of researchers. Nonetheless, it is imperative to ensure that tests' validity and reliability are not reduced relative to the originals.

Because the $M-$measures in general, including the CMT and NMT, are unique tools for cognitive assessment in developmental research, clinical settings, and neuroscience, shorter measures should be assets. In addition, shorter tests could lead to increased use when time is saved. However, it is of paramount importance that the accuracy and reliability of the original tests be maintained. Smith, McCarthy, and Anderson (2000) outlined methodological caveats, and they offered methodological steps for short-form development, which we can use. Theoretically, we anticipate that classic, recent, and new measures of mental attentional capacity will document developmental trends across childhood and adolescence. We expect that tasks with higher interference (i.e., more misleading) will better assess cognitive patterns and that task domain will not elicit different performance but instead equivalent performance across age groups. A practical aim of the current study was to validate the NMT and to examine whether the reliability and validity of the CMT and NMT can be preserved within abbreviated tasks.

Method

Participants

In total, 483 participants were tested in the following six age groups: Grade 2 (7–8 years), Grade 4 (9–10 years), Grade 6 (11–12 years), Grade 8 (13–14 years), Grade 10 (15–16 years), and adults (19–30 years), comprising 209 male and 274 female participants. Table 1 shows the distribution of participants by age group and gender. Groups A and B in the table describe the characteristics of two groups randomly divided from the larger group to conduct analyses (see "Analyses" section for details on random group assignment).

Of the adult sample (n = 226), 101 adults were recruited from the community and 125 were recruited from the university and received course credit. Children were recruited from four public schools. All participants or their parents signed an informed consent form agreeing to participate. Children received a small stationery gift (e.g., pens and pencils) for their participation in each session. The study received ethics approval.

Materials

Color and number matching tasks

Two tasks were completed by all participants: a CMT (Arsalidou et al., 2010) and an NMT (current study; see Fig. 1). Participants were asked to indicate whether the current image contained the same colors for the CMT or the same numbers for the NMT as the previous one by pressing a button.

Table 1
Participant characteristics.

Group	*n*	Male	Age (years)	
			M	*SD*
Grade 2	62	34	7.26	0.44
A	31	15	7.23	0.43
B	31	19	7.29	0.46
Grade 4	48	20	9.04	0.2
A	24	8	9.00	0
B	24	12	9.08	0.28
Grade 6	61	28	11.11	0.32
A	31	15	11.06	0.25
B	30	13	11.17	0.38
Grade 8	37	12	13.37	0.45
A	18	6	13.22	0.43
B	19	6	13.32	0.48
Grade 10	49	15	15.22	0.47
A	25	8	15.2	0.41
B	24	7	15.25	0.53
Adults	226	100	22.19	2.52
A	113	50	21.89	2.39
B	113	50	22.48	2.63
Total	483	209		
A	242	102		
B	241	107		

Within both tasks, there are both low-interference and corresponding high-interference versions. In the CMT, low interference (i.e., facilitating contexts) is assessed using the Balloon task (CMT-B) and high interference is assessed using the Clown task (CMT-C). In both tasks, a picture of either balloons (CMT-B) or a clown (CMT-C) is presented with six levels that are manipulated by the number of relevant colors present (purple, orange, yellow, pink, brown, gray, and red). The location and repetition of relevant colors and the colors blue and green are irrelevant cues that must be ignored in both tasks. In the CMT-C, the prominent figure of the clown itself, including the colors in the clown's face, is an additional irrelevant cue to be ignored, resulting in a higher misleading complexity measure due to increased interference.

In the NMT, low interference is assessed using the Squares task (NMT-S) and high interference is assessed using the Fours task (CMT-F). In both tasks, a picture of numbers arranged in either a square pattern (NMT-S) or a large number four pattern (NMT-F) is presented with six levels that are manipulated by the number of relevant digits present (i.e., 1, 2, 3, 4, 6, 7, and 8). The location and repetition of relevant numbers and the numbers 0 and 5 are irrelevant cues that must be ignored in both tasks. In

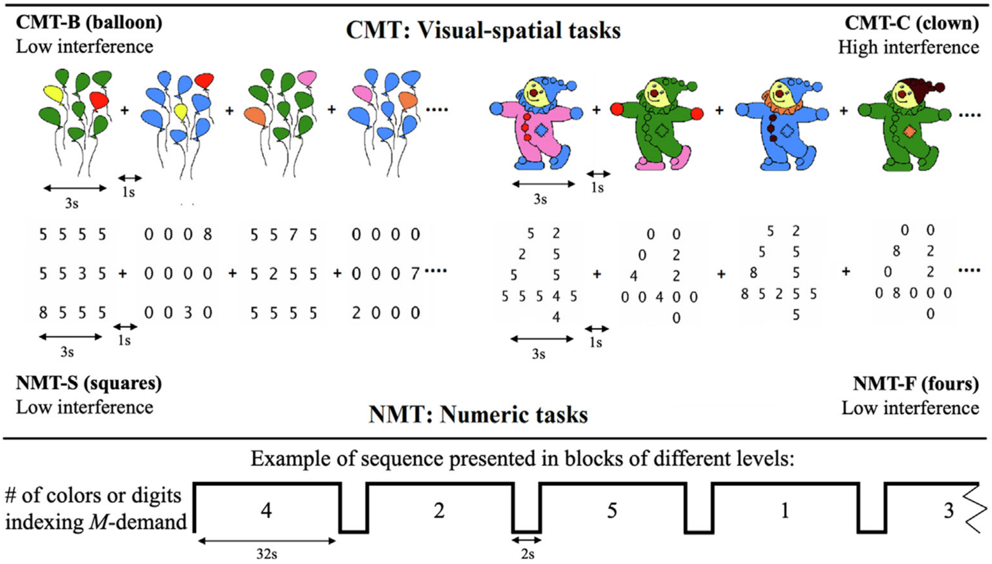

Fig. 1. Examples of Color Matching Task (CMT) and Number Matching Task (NMT) images and sequence of presentation. Participants were asked to indicate whether the current image contained the same colors for the CMT or numbers for the NMT as the previous one by pressing a button. We show mental demand level 3 for low interference (CMT-B and NMT-S) and mental demand level 4 for high interference (CMT-C and NMT-F). Each stimulus was presented for 3 s with a 1-s interstimulus interval. A trial was constituted by consecutive stimuli (e.g., second picture was compared with first picture, third picture was compared with second picture). A block was constituted by 7 trials of the same difficulty. Blocks were separated by a 2-s interblock stimulus. The tasks were administered in two runs. Each run included two blocks for every difficulty level in a pseudorandom order.

the NMT-F, the salient figure of the large number four itself is an additional irrelevant cue to be ignored, resulting in a higher-complexity measure due to interference. Irrelevant cues require participants to inhibit impulsive responses and ignore these aspects of the stimuli.

Using a match to sample procedure, participants were asked to indicate whether the relevant items (digits or colors) in the current stimulus match those of the previously presented stimulus while disregarding irrelevant features. Both figurative schemes (e.g., colors or numbers) and executive schemes (i.e., higher-order operative schemes such as task rules) that are not facilitated by the situation count toward the estimate of mental (*M*-) demand of the task. Specifically, *M*−demand is increased by the number of relevant features (digits or colors; $n = 1$–6) in each stimulus, whereas the executive demand (i.e., goals) of the task remains constant, producing parametrically scaled levels of *M*−demand (e.g., number of colors plus executive demand). The amount of mental attention required to complete each task is indexed by both the number of relevant features (colors or digits; $n = 1$–6) and the amount of interference ($n + 1$ or $n + 2$). The amount of interference in these tasks creates either a facilitating context (low interference) or a misleading context (high interference). Task analyses estimate that the facilitating task requires one constant executive scheme of identifying and matching the relevant colors, whereas the misleading task requires an additional constant executive scheme (i.e., higher-order operative scheme), which is related to effortfully extracting relevant colors from an embedding context by ignoring irrelevant features (Arsalidou et al., 2010). Thus, an *M*−demand of $n + 1$ is associated with low-interference tasks (CMT-B and NMT-S), and that of $n + 2$ is associated with high-interference tasks (CMT-C and NMT-F). The addition of irrelevant cues in the high-complexity tasks required participants to use more mental attention, which causes (our task analyses show; Arsalidou et al., 2010) one additional unit to be added, resulting in $n + 2$. Thus, one unit shared by the two versions of the task related to the mental operation reflects identifying and matching colors, and the second unit related to the mental operation of extracting relevant features from an embedded context is shared only by the misleading versions of the two tasks. The score of *M*−capacity is indexed by the highest (*M*−demand)

Table 2
Participant characteristics for Figural Intersection Task.

Group	*n*	Male	Age (years)	
			M	*SD*
Grade 2	31	18	7.29	0.46
Grade 4	35	15	9.03	0.17
Grade 6	40	18	11.05	0.22
Grade 8	26	10	13.19	0.40
Grade 10	31	7	15.26	0.51
Adults	148	66	22.05	2.45
Total	311	134		

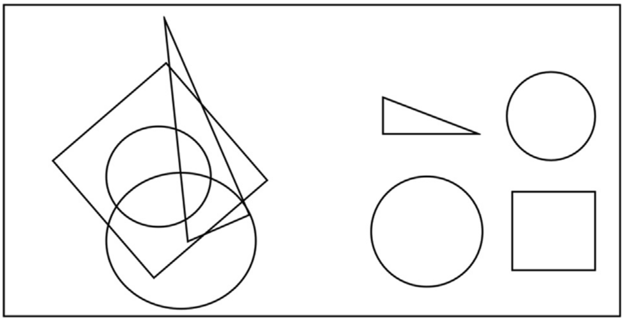

Fig. 2. An example of difficulty level 4 in the Figural Intersection Task (FIT).

level participants can pass reliably with at least 70%, allowing for only one such percentage level to drop below.

Figural Intersection task

The Figural Intersection Task (FIT; Pascual-Leone & Baillargeon, 1994) is a classic measure of *M*—capacity, with scores on the FIT having been shown to agree with the theoretical predictions of *M*—capacity in many studies (e.g., Johnson et al., 2003; Pascual-Leone, 1970; Pascual-Leone & Baillargeon, 1994; Pascual-Leone & Ijaz, 1989; Pascual-Leone et al., 2000; see Arsalidou & Im-Bolter, 2017, and Onwumere & Reid, 2008, for reviews). In the current study, the FIT was used as one comparison measure to evaluate the validity and reliability of the NMT. The FIT was not completed by all participants due to school absences on the days of testing (see Table 2).

The FIT (see Fig. 2) is a visual–spatial task in which participants are asked to attend to the shapes on the right and find the common overlap (i.e., point of total intersection) of all those shapes on the left. The size, location, and rotation of the shapes, as well as extra shapes, are irrelevant cues and must be ignored. Levels of difficulty are parametrically scaled by progressively increasing the number of shapes (2–8). A total of 28 items are to be completed (4 items with seven levels of difficulty), and *M*—capacity is assessed by the highest difficulty level participants can pass reliably at least 70% of the time, allowing for only one such percentage level to drop below.

Procedure

Color and number matching tasks

Participants were tested in a quiet room where a laptop computer was set up. The CMT and NMT were administered in a randomized order on different days after participants had training for each task. Children and adults received the same training on a computer. After being greeted by the researcher, participants were familiarized with the stimuli and rules of the task using scripted instructions and a PowerPoint presentation. First, participants were asked to name colors (those involved in the CMT) or numbers (those involved in the NMT). Next, using PowerPoint slides, the stimuli were presented as examples and the experimenter explained the rules of the task, such as the rule that blue and green colors or the numbers 0 and 5 are not important (i.e., irrelevant) when making decisions. In addition, participants learned that relevant colors (in the CMT) or numbers (in the NMT) can change locations and to ignore the face of the clown for the CMT or the large number for the NMT. Then, participants completed a short practice with 12 trials, with low levels of difficulty (i.e., n = 1–3 colors or numbers), to get acquainted with the pace and timing of the game. Next, participants were asked to complete the real task on their own on a computer. This task included all difficulty levels (n = 1–6 colors or numbers), which were presented in a pseudorandom order. Participants responded by pressing the period (.) for same and a slash (/) for different on a standard keyboard. Each stimulus was presented for 3 s with a 1-s interstimulus interval; a short 2-s interval separated the difficulty blocks, which were pseudorandomly presented (e.g., number of relevant features: 3, 5, 1, 6, 2, 4), for both the training and experimental sessions.

Figural Intersection task

The FIT was administered either individually for adult participants or in groups of 5 to 10 for children in schools. Instructions were scripted, and researchers provided instructions using the whiteboard in classroom settings. First, the children and adults were familiarized with the idea of putting dots inside the shapes on the right side of the page and finding the same shapes on the left. Then, participants learned that they needed to place only one dot in all the shapes on the left side of the page at the same time (i.e., point of total intersection). Next, participants learned that shapes could be different sizes (e.g., a small circle on the right and a big circle on the left) but were still considered as being the same shape. They also learned that shapes could rotate but were still considered as being the same shapes. Lastly, participants learned that shapes that were found on the left and were not found on the right were not important (i.e., irrelevant) and should be ignored. Participants practiced these steps in their booklets using a red pen. Feedback was provided to ensure that participants understood the task and instructions.

Analyses

To identify significant agreement among test scores, the majority of the analyses conducted in the current study used equivalence tests (e.g., Mara & Cribbie, 2018; Wellek 2003). Traditional t tests, which test whether two samples are different, pose limitations for researchers who would like to assess whether sample means are significantly equal rather than significantly different. Similar to t tests, equivalence tests can be used to draw statistical inferences based on applying a threshold on observed data. In equivalence tests, the null hypothesis states that there is a difference (i.e., outside the specified equivalence bounds), whereas the alternative hypothesis is an effect that is within the equivalence bounds, suggesting that two means are significantly equivalent. Thus, if a t test shows that two means are not significantly different, it is unclear whether the means are in fact statistically equivalent because they can be neither statistically different nor statistically equivalent. A test of statistical equivalence, such as the Two One-Sided Test (TOST), provides an inferential statistic for identifying comparisons that yield significantly similar results and serves as a more stringent test for claiming that two means are comparable.

The current study used the TOST procedure to determine whether scores across tasks were equivalent. The null hypothesis for the TOST is that there is a difference in variance that falls outside of the expected equivalence bounds, and the alternative hypothesis is that the difference in variance will fall inside the expected equivalence interval (Mara & Cribbie, 2018). An expected mean difference of 1 was used to conduct all analyses. The lower bound was set to -1 and the upper bound was set to + 1. If the null hypothesis is rejected, we can conclude that the difference between the two means falls within the equivalence bounds (i.e., greater than -1 and less than + 1) and is statistically equivalent. Equivalence bounds were set based on previous studies that tested equivalence of the CMT and other mental attentional capacity measures (Arsalidou et al., 2010; Powell et al., 2014). In addition, as can be

Table 3
Means and standard deviations for scores on tasks across age groups.

	CMT-B		CMT-C		NMT-S		NMT-F		FIT	
Age (years)	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
7–8	3.76	1.63	3.34	1.24	2.97	1.34	3.18	1.25	3.39	1.43
9–10	4.42	1.37	3.90	1.37	3.31	1.19	3.46	1.17	4.74	1.50
11–12	5.74	1.33	5.10	1.45	4.52	1.32	4.69	1.29	5.08	1.80
13–14	6.05	1.25	6.38	1.30	5.51	1.12	5.19	1.29	6.00	1.55
15–16	5.65	1.55	5.96	1.81	5.53	1.37	5.47	1.68	5.94	1.81
19–30	6.30	0.89	6.33	1.36	5.94	1.00	6.35	1.50	5.80	1.44

Note. CMT, Color Matching Task (B, balloon; C, clown); NMT, Number Matching Task (S, Squares; F, Fours); FIT, Figural Intersection Task.

seen in Table 3, the standard deviations and range for scores on individual tasks all are greater than 1 except the standard deviation for adults on the CMT-B. We expect that this degree of variability within each task reflects the appropriate decision to expect a mean difference of 1 across tasks.

Equivalence analyses were performed in R Version 3.4.3 (R Core Team, 2017) using the package “equivalence” (Robinson, 2016) and the TOST function. Both *t* test and correlation analyses were performed using SPSS software Version 23 (IBM Corp., Armonk, NY, USA). Random assignment of groups was completed using Microsoft Excel. To preserve variability in the sample and consider as many data points as possible, no outliers were removed. For reference, we estimated the number of over- performers and under-performers using the interquartile range (IQR) method (i.e., 1.5 IQR below quartile 1 or more than 1.5 IQR above quartile 3 (see Table S1 in online supplementary material).

Preprocessing: Group assignment

One aim of the current study was to determine whether a reliable *M*−capacity score can be obtained with a reduction in the number of blocks and by extension the number of trials needed. We note that all children completed the full version of all tasks. Group assignment was performed to identify whether fewer trials yielded comparable scores as full versions of the task. Thus, for data analyses, the full score of one group was compared with the partial scores of another group. The CMT and NMT follow the same protocol design. Each task block had 7 trials (i.e., 8 stimuli) corresponding to a difficulty level. Each block was pseudorandomly presented two times (i.e., 14 trials) in each run, and we had two runs (i.e., a total of 28 trials per difficulty level). The task was designed to have two runs to provide a short break for children in between runs. Each run took about 7 min to complete. Specifically, during 7 min children completed two blocks for each difficulty level (i.e., 14 trials), during 14 min children complete four blocks for each difficulty level (i.e., 28 trials), and intermediately, if we consider half the trials in the second run (7 + 3.5 min = 10.5 min), children completed three blocks for each difficulty level (i.e., 21 trials). This timing and trials were the same for each version of the CMT and NMT. Consequently, the full task with 28 trials took 28 min, abbreviated versions with 21 trials took 21 min, and 14 trials took 14 min to complete.

To test the hypothesis on whether the full test could yield comparable results to the abbreviated version of the task, each age group in the original sample was randomly divided into two groups: Group A and Group B (see Table 1). Group A and Group B completed the full version of the tasks; however, in the statistical analyses, we evaluated whether their performance would be the same if we considered a partial number of trials for one of the groups. Random assignment to separate groups was done to ensure that when conducting equivalence analyses there was no bias by having the same scores from one block included in one of the different block combinations. For example, participants’ score for Block 1 to Block 2 would be included in their score for Blocks 1 to 4, creating a bias for equivalence.

Random allocation to Group A or Group B was established using the following steps. First, a random number was generated using the = *RAND()* function in Microsoft Excel for each participant. Second, the randomly generated numbers were then sorted from smallest to largest, shuffling participants’ current

Table 4
Results among CMT, NMT, FIT, and predicted $M-$capacity.

CMT versions									
Age (years)	Pred.	vs.	CMT-B	vs.	CMT-C	vs.	FIT	vs.	Pred.
7–8	3	≠	3.76	=	3.34	=	3.39	=	3
9–10	4	=	4.42	=	3.90	≠	4.74	≠	4
11–12	5	≠	5.74	=	5.10	=	5.08	=	5
13–14	6	=	6.05	=	6.38	≠	6.00	=	6
15–16	7	≠	5.65	=	5.96	=	5.94	≠	7
19–30	7	=	6.30	=	6.33	=	5.80	≠	7
Misleading versions									
Age (years)	Pred.	vs.	CMT-C	vs.	NMT-F	vs.	FIT		
7–8	3	=	3.34	=	3.18	=	3.39		
9–10	4	=	3.90	=	3.46	≠	4.74		
11–12	5	=	5.10	=	4.69	=	5.08		
13–14	6	=	6.38	≠	5.19	≠	6.00		
15–16	7	≠	5.96	≠	5.47	≠	5.94		
19–30	7	=	6.33	=	6.35	=	5.80		
Facilitating versions									
Age (years)	Pred.	vs.	NMT-S	vs.	CMT-B	vs.	FIT		
7–8	3	=	2.97	≠	3.76	=	3.39		
9–10	4	=	3.31	≠	4.42	=	4.74		
11–12	5	=	4.52	≠	5.74	≠	5.08		
13–14	6	=	5.51	=	6.05	=	6.00		
15–16	7	≠	5.53	=	5.65	≠	5.94		
19–30	7	≠	5.94	=	6.30	=	5.80		
NMT versions									
Age (years)	Pred.	vs.	NMT-F	vs.	NMT-S	vs.	FIT		
7–8	3	=	3.18	=	2.97	≠	3.39		
9–10	4	=	3.46	=	3.31	≠	4.74		
11–12	5	=	4.69	=	4.52	≠	5.08		
13–14	6	≠	5.19	=	5.51	=	6.00		
15–16	7	≠	5.47	=	5.53	≠	5.94		
19–30	7	=	6.35	=	5.94	=	5.80		

Note. $p < .05$. Expected mean difference of 1 was used to calculate the equivalence intervals. Equal sign (=) signifies equivalence; not equal sign (≠) signifies nonequivalence. Scores represent the mean of the group. In the first section of the table, we are comparing the Color Matching Task (CMT) versions (B, balloon; C, clown). In the second section, we are comparing misleading versions. In the third section, we are comparing facilitating versions. In the fourth section, we are comparing Number Matching Task (NMT) versions (F, Fours; S. Squares). Pred., predicted $M-$capacity; FIT, Figural Intersection Task.

placement in the spreadsheet. Third, the first half of participants were assigned to Group A and the second half were assigned to Group B, resulting in equal samples when possible. Group A scores were computed from all four blocks (1 to 4, 28 trials). Group B was the comparison group. Scores for Group B were computed from Blocks 1 and 2 (1 to 2, 14 trials) and Blocks 1, 2, and 3 (1 to 3, 21 trials), depending on the comparison being made. Two-sample t tests revealed that Group A and Group B were not significantly different on age and $M-$scores (for the complete task, 28 trials) across all tasks and age groups (see Table S2 in supplementary material).

Results

Number matching tasks

Paired TOSTs were conducted to examine the validity of the NMT by comparing mean scores among the CMT, FIT, and theoretical predictions of $M-$capacity (Table 4 and Fig. 3). Table 4 was designed to compare equivalence among means of the CMT, predicted scores, and FIT as reported by Arsalidou et al. (2010). The second and third sections of Table 4 compare the misleading and facilitating versions of the CMT and NMT with one another, and the fourth section compares the NMT versions with predicted scores and the FIT. Scores on the NMT-S and CMT-B were equivalent for half the

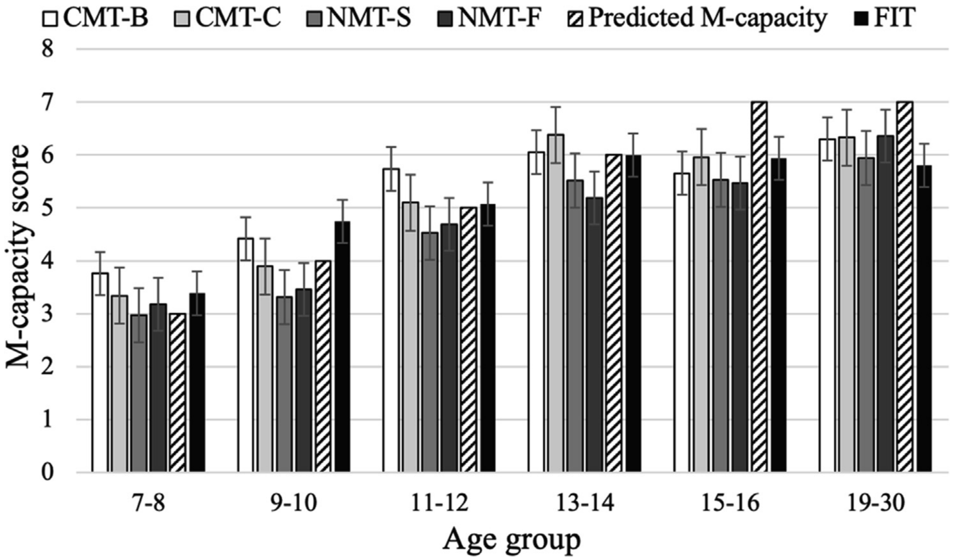

Fig. 3. Mean scores as a function of age. Error bars represent standard errors (±1). Diagonal stripes represent the predicted *M*−capacity for each age group. CMT, Color Matching Task (B, balloon; C, clown); NMT, Number Matching Task (S, Squares; F, Fours); FIT, Figural Intersection Task.

Table 5
Significant correlations among tasks, age, and predicted *M*−capacity.

	CMT-B	CMT-C	NMT-S	NMT-F	FIT	PMC
Age	.49**	.51**	.62**	.61**	.34**	.89**
CMT-B	–	.65**	.65**	.57**	.38**	.56**
CMT-C	.56**	–	.67**	.59**	.37**	.61**
NMT-S	.55**	.54**	–	.73**	.40**	.69**
NMT-F	.38**	.41**	.59**	–	.37**	.64**
FIT	.27**	.25**	.27**	.23**	–	.42**
PMC	.31**	.38**	.41**	.26**	.28**	–

Note. Bivariate Pearson's *r* correlations are above the diagonal. Partial Pearson's *r* correlations controlling for age are below the diagonal. All results were significant at ***p* <.01 (two-tailed). CMT, Color Matching Task (B, Balloon; C, Clown); NMT, Number Matching Task (S, Squares; F, Fours); FIT, Figural Intersection Task; PMC, predicted *M*−capacity.

age groups. Scores on the NMT-F and CMT-C were equivalent for all age groups except the two adolescent groups. Both the NMT-S and NMT-F were equivalent to the theoretically predicted *M*−scores for all age groups except the two adolescent groups, which may have lower scores due to motivational issues. Consistent with previous research (Arsalidou et al., 2010; Powell et al., 2014) and with theoretical predictions (Pascual-Leone et al., 2000), high-interference tasks (CMT-C and NMT-F) showed better equivalence with the predicted *M*−capacity and FIT.

Correlational methods were used to examine developmental performance patterns and relations among tasks. All tasks were highly correlated with each other, predicted *M*−capacity and chronological age, and remained significant (*p* <.01, two-tailed) when the variance for age was removed (see Table 5). High correlations for the CMT-B and NMT-S, as well as for the CMT-C and NMT-F, are suggestive of concurrent construct validity.

Paired TOSTs were conducted between scores on Block 1 to Block 2 (first half of test; 14 trials) and scores on Block 3 to Block 4 (second half of test; 14 trials) for each task to examine the consistency of scores (Table 6). The first and second halves of the test were equivalent for all groups and tasks except

Table 6
Comparison of scores on Block 1 to Block 2 versus Block 3 to Block 4 for the whole sample (i.e., Groups A and B combined).

	CMT-B			CMT-C			NMT-S			NMT-F		
Age (years)	1 to 2	vs.	3 to 4	1 to 2	vs.	3 to 4	1 to 2	vs.	3 to 4	1 to 2	vs.	3 to 4
7–8	4.26	=	3.88	3.85	=	3.44	3.19	=	3.26	3.45	=	3.35
9–10	4.83	=	4.54	4.67	=	4.13	3.71	=	3.69	3.71	=	3.98
11–12	6.03	=	5.82	5.69	=	5.21	4.60	=	5.02	5.23	=	4.85
13–14	6.32	=	6.02	6.76	≠	5.92	5.78	=	5.45	5.51	=	5.54
15–16	5.98	=	5.88	5.96	=	6.10	5.59	=	5.71	5.71	=	5.86
19–30	6.38	=	6.39	6.55	=	6.47	5.99	=	6.11	6.65	=	6.44

Note. Scores represent the *M*–score of each age group for Block 1 to Block 2 and for Block 3 to Block 4. An expected mean difference of 1 was used (p <.05). CMT, Color Matching Task (B, Balloon; C, Clown); NMT, Number Matching Task (S, Squares; F, Fours).

the Grade 10 sample (15–16 years) on the CMT-C. Paired TOSTs between scores on each block (i.e., 1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4, and 3 vs. 4; 7 trials per block) were also conducted to assess the reliability of each task (see Table S3 in supplementary material).

Independent t tests were conducted to see whether there were any differences between genders per task and age group. Significant differences were found between male and female participants in Grade 10 on the CMT-C, $t(47) = -3.03$, p =.004, and adults on the NMT-F, $t(224) = 0.023$. Female participants in these two groups had higher means on the CMT-C (female: $M = 6.44$; male: $M = 4.87$) and NMT-F (female: $M = 6.56$; male: $M = 6.10$), and female participants in Grade 2 had higher means on the CMT-C, $t(60) = -2.005$, p =.049 (female: $M = 3.68$; male: $M = 3.06$). No gender differences were found for the other tasks across age groups (see Table S4 in supplementary material).

Abbreviations of CMT and NMT

Independent-sample TOSTs were conducted to compare the scores on all four blocks (1 to 4, 28 trials; Group A) with those on the other two block combinations (1 to 2, 14 trials, and 1 to 3, 21 trials; Group B). Scores using half the tasks were very similar to scores from the full task (Table 7 and Fig. 4). For low-interference tasks (CMT-B and NMT-S), scores on the full test (Blocks 1 to 4) were equivalent for all age groups when compared with scores from half the test (Block 1 to Block 2) except Grade 2.

Table 7
Results of equivalence tests comparing all four blocks (1 to 4) with the other block combinations (1 to 2 and 1 to 3).

	CMT-B					CMT-C				
Age (years)	1 to 4	vs.	1 to 2	vs.	1 to 3	1 to 4	vs.	1 to 2	vs.	1 to 3
7–8	3.58	≠	4.23	≠	3.94	3.16	≠	3.97	≠	3.67
9–10	4.63	=	4.58	≠	4.29	4.13	≠	4.54	=	3.92
11–12	5.84	=	6.03	≠	5.77	5.19	≠	5.73	=	5.03
13–14	6.33	=	6.21	=	5.68	6.56	=	6.63	=	6.32
15–16	5.92	=	5.83	≠	5.63	5.80	≠	6.00	=	5.79
19–30	6.27	=	6.40	=	6.42	6.31	=	6.57	=	6.45

	NMT-S					NMT-F				
Age (years)	1 to 4	vs.	1 to 2	vs.	1 to 3	1 to 4	vs.	1 to 2	vs.	1 to 3
7–8	2.71	≠	3.45	≠	3.29	3.00	≠	3.51	=	3.32
9–10	3.42	=	3.46	=	3.25	3.33	=	3.67	=	3.33
11–12	4.42	=	4.60	=	4.67	4.94	=	4.90	=	4.73
13–14	5.67	=	5.63	=	6.63	5.17	≠	5.47	=	5.32
15–16	5.60	=	5.46	=	5.42	5.40	≠	5.92	≠	5.96
19–30	5.92	=	5.95	=	5.96	6.32	=	6.62	=	6.58

Note. Scores represent the mean of each group. Data from Group A were used to generate scores for all blocks (1 to 4), and scores from Group B were used to generate scores for Block 1 to Block 2 and for Block 1 to Block 3. An expected mean difference of 1 was used (p <.05). CMT, Color Matching Task (B, Balloon; C, Clown); NMT, Number Matching Task (S, Squares; F, Fours).

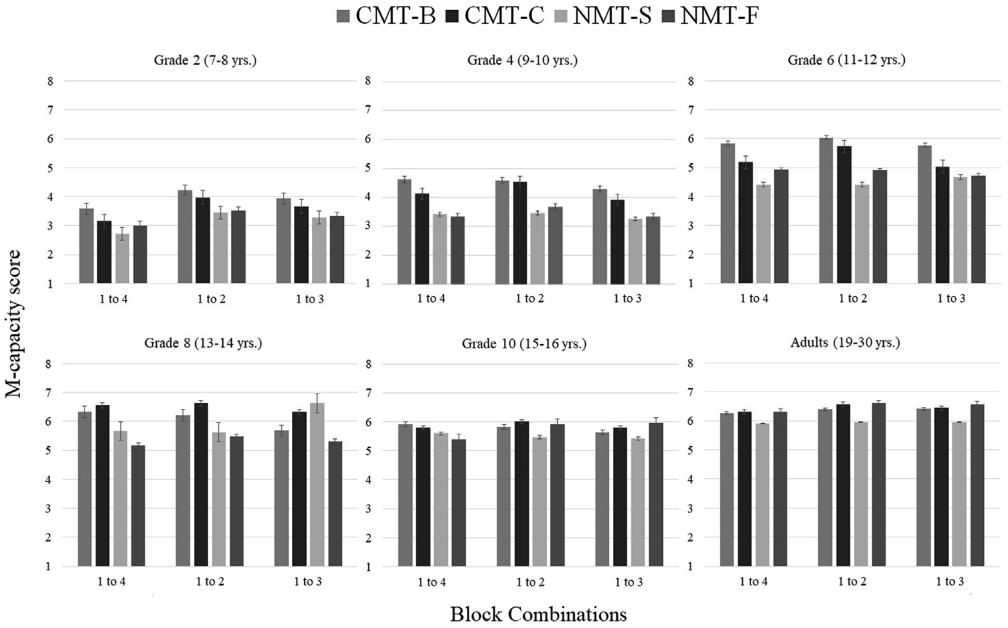

Fig. 4. *M*−capacity scores per block combination for each age group. Error bars represent standard errors (±1). CMT, Color Matching Task (B, balloon; C, clown); NMT, Number Matching Task (S, Squares; F, Fours).

For high-interference tasks (CMT-C and NMT-F), scores on the full test were equivalent for all age groups when compared with scores using only three blocks of the test (1 to 3, 21 trials) except Grade 2 for the CMT-C and Grade 10 for the NMT-F. Complementary analyses with more lenient equivalence bounds (i.e., 1.2 and 1.5) show that all comparisons reached equivalence (see Tables S5–S12 in supplementary material).

Discussion

The purpose of this study was to evaluate effects of age, task domain, task interference, and task difficulty (*M*−demand) on performance with measures of mental attentional capacity and to determine the validity and reliability of the full and abbreviated forms of *M*−capacity tasks. Results showed increases in mental attentional capacity across domains, supporting the notion of domain-general trends across development. Tasks with higher interference that introduced misleading context during problem solving improved assessment of these stages, consistent with previous findings in children (Arsalidou et al., 2010) and adults (Engle, 2001). Practically, full and abbreviated versions of the CMT and NMT could be used in future research and practice to assess mental attentional capacity starting in school-aged children of 7 years.

The two variants of the CMT and NMT differ only in contextual information, such that the higher misleading complexity tasks in fact produce increased interference due to irrelevant cues. In the CMT-C and NMT-F, the global very salient figure in which the colors or numbers are embedded is completely irrelevant to the task at hand. According to the TCO, to extract task-relevant cues (colors or numbers) for processing, the irrelevant cues need to be inhibited, which causes the task to be more attentionally demanding (Pascual-Leone, 1970, 1989; Pascual-Leone & Johnson, 2005, 2011). Participants tend to score slightly higher on the CMT compared with the NMT, and this effect is more evident in the facilitating versions of the task (CMT-B and NMT-S). This may be because the colors are more salient, whereas the numbers are made up of lines and curves that need a little bit more processing to be deciphered. In fact, we observed that child groups over-performed on the CMT-B, replicating past findings (e.g., Arsalidou et al., 2010) and resulting in nonequivalent results when compared with the NMT-S. However, results of adolescents and adults showed significant equivalence among scores of the CMT-B and NMT-S. Our data are consistent with previous research with the CMT showing that the higher-interference (i.e., misleading complexity) task variants are a better measure of $M-$capacity (Arsalidou et al., 2010; Powell et al., 2014).

Because the CMT-C is misleading (i.e., it contains more irrelevant features), it was expected to yield greater equivalence with the theoretically predicted $M-$capacities across age groups compared with the CMT-B, which showed equivalence for half the age groups (Table 4). Younger children, but not teenagers and young adults, tended to score higher on the CMT-B than on the CMT-C, in agreement with previous results (Arsalidou et al., 2010; Powell et al., 2014). Notably, when nonequivalence is observed in children, it is often because of over-performance, whereas when nonequivalence is observed in adolescents and young adults, it is often because of under-performance. This may be associated with a lack of engagement and motivational issues reflected in the performance of these age groups (Goss & Sonnemenn, 2017; Mann, 2018; Martin, 2009). Further research is needed to understand the relation between motivation and cognitive abilities.

Motivational processes are said to affect performance on cognitive tasks because individuals may lack the goal to perform well and avoid negative judgments about their competence (Dweck, 1986). Significant gender differences were found for the Grade 10 sample on the CMT-C and CMT-F (the two higher-interference variants), with female participants scoring higher. This is interesting because some research has shown that male participants perform better on visual–spatial and verbal working memory tasks (Geiger & Litwiller, 2005). However, there are inconsistent findings in that other research shows that female participants perform better on both (Duff & Hampson, 2001). Alternatively, these results may be due to the lack of engagement shown by males in secondary education; research has shown that levels of achievement and engagement in males have been on the decline (Martin, 2004; Weaver-Hightower, 2003). Males are more disengaged in class, are more unwilling to do additional work, and have a negative attitude toward school (Mann, 2018; Martin, 2007). Future research should consider more age-appropriate motivators for adolescents. Overall, with the exception that female participants in Grade 10 over-performed compared with male participants in Grade 10, our results suggest a lack of gender differences in $M-$capacity (the maturational component of working memory) in both visual–spatial and numeric domains.

Scores on the NMT-S and NMT-F yielded equivalent results with all age groups. When each was compared with theoretical scores (i.e., predicted mental attentional capacity), equivalence was observed among all scores except for two groups in either teenage participants or young adults (Table 4). This suggests that the NMT-S, unlike the CMT-B, is a stricter measure of $M-$capacity with less opportunity to over-perform than we initially expected given that the identity of each color is made salient by the separate balloons, whereas the NMT-S does not share this pop-out phenomenon. However, the NMT-F showed more equivalence with the FIT, considered to be a high-interference task due to irrelevant cues encountered when seeking the figural intersection, suggesting that the NMT-F is a better measure of $M-$capacity than the NMT-S, which had equivalence for only two of the six groups. Future research is needed to further investigate the tendency of the NMT to slightly under-estimate participants' $M-$capacity. Overall, equivalence tests discussed above generally showed performance on these measures closely following theoretically predicted increases across development (Arsalidou et al., 2010; Pascual-Leone, 1970; Pascual-Leone & Baillargeon, 1994; Pascual-Leone & Johnson, 2005, 2011).

Correlation analyses produced results that are in agreement with the claim that the number of information units that can be simultaneously held (and mentally manipulated) increases with age (Table 5). We note the high correlations for the CMT-B and NMT-S as well as for the CMT-C and NMT-F. Expectedly, performance significantly improved as a function of age. However, age is not the only factor that could explain scores because correlations remain significant even when the variance for age is removed. Speculating on the factors that may explain the significant correlations after controlling for age may relate to genetic and environmental factors. Moreover, individual differences in $M-$capacity between children of the same age may relate to maturational factors. A study on giftedness reported that gifted children have an $M-$capacity of about one unit more than their peers on average (Johnson et al., 2003). We also speculate that it may also relate to method variance. Method variance reflects characteristics of the task; although at different domains, all tasks followed a similar task construction (e.g., constant rules, multiple levels) that also shared timing and $M-$demand order. This is more evident for computerized tasks that share timing and presentation style. Partial correlations between computerized tasks and the FIT showed comparable scores to past studies that used $M-$capacity tasks that differed in their general administration practice and reported significant partial correlations of about .25 in two samples of school-age children (Morra, 1994), a result that was replicated with preschoolers (Panesi & Morra, 2020) and adults (Morra, Camba, Calvini, & Bracco, 2013), although higher partial correlations were also reported (Bisagno & Morra, 2018). Future research examining neurophysiological factors associated with mental attention will be useful for identifying marker variables that account for this shared variance. Overall, correlational results (among the tasks and with the predicted $M-$scores) add construct validity to the method of mental attention measurement in general and to the new numeric $M-$capacity task; these all are $M-$capacity measures across distinct content domains.

Concordance of scores across different blocks and block combinations shows the task's reliability. Reproducible scores between different blocks for all tasks showed that participants' performance was reliable throughout the task and suggests that the CMT and NMT are reliable measures of $M-$capacity within a misleading context (a context that requires attentional inhibition). Thus, in agreement with previous results in children (Arsalidou et al., 2010; Powell et al., 2014) and adults (e.g., Engle & Kane, 2004), the misleading version of the NMT (i.e., NMT-F) was shown to be, as predicted, a better measure of mental attentional capacity.

Support for abbreviated measures

After confirming the validity of the NMT, equivalence analyses were performed to assess whether using fewer trials of the CMT and NMT would provide a valid measure of $M-$capacity. A group of participants (Group A) whose scores were from the whole test (Blocks 1 to 4) were compared with a different group of participants (Group B) whose scores were from different combinations of blocks (1 to 2 and 1 to 3). Results showed across all age groups that for both low-interference tasks (CMT-B and NMT-S), scores from the whole task were equivalent to the half-test combination (Block 1 to Block 2), except for Grade 2 (7–8 years) on both tasks (Table 7). Scores on full versions (Blocks 1 to 4) of the high-interference tasks were equivalent to scores from Blocks 1, 2, and 3 (1 to 3) for all age groups except Grade 2 (7–8 years) on the CMT-C and Grade 10 (15–16 years) on the NMT-F. Many of the mean differences between scores from Blocks 1 to 4 and those from the other two block combinations (1 to 2 and 1 to 3) were less than 0.50 (i.e., half a point). In addition, two-sample t tests revealed that there were no significant differences (p <.05) between scores on the full tasks when compared with those on the other block combinations for all tasks and age groups except Grade 2 (7–8 years) on the CMT-C for Block 1 to Block 2 (see Tables S13 and S14 in supplementary material for results). By observing the means, some of the tasks and age groups showed scores closer to their theoretically predicted $M-$capacity on the abbreviated tasks compared with the full tasks. Grade 10 participants in Group B performed better when scores were from an abbreviated task than Group A, whose scores were from the full test. For example, on the NMT-F the mean for Group A (Blocks 1 to 4) was 5.40, whereas the mean for Group B in Blocks 1 and 2 was 5.92, which is closer to the predicted $M-$capacity of 7 for this age group. These results suggest a valid $M-$capacity score using half the task.

Conclusions

This study investigated properties of mental attentional capacity in children, adolescents, and adults. Findings suggest a common trajectory of development as a function of domain. Task interference also affects scores differently by domain and by age. For instance, in color tasks younger children benefited from low interference. This study determined validity of the new numeric $M-$capacity task and verified theoretical predictions of the developmental increase in $M-$capacity. Specifically, findings on the CMT were replicated (Arsalidou et al., 2010; Powell et al., 2014); and it was shown that the higher misleading complexity variant of the NMT (with more interference) improved the $M-$capacity assessment across development. In addition, processing visual–spatial and numeric information was shown to be more efficient with growing age. This study determined that an abbreviated versions of the CMT and NMT can produce valid and reliable $M-$scores. Because this is the first study to report results on the numeric measure of mental attentional capacity, more studies are needed to replicate and confirm developmental patterns. Theoretically, our findings contribute to knowledge on psychological processes that underlie cognitive development in children, adolescents, and adults. Practically, developmental scores can serve as a benchmark for comparison for future studies with children with typical and atypical development. In addition, abbreviated tasks will be an asset for future research because they can save researchers time and money when conducting both behavioral and neuroimaging studies with children and adults. Overall, findings and methods presented in this work can benefit understanding and future research in psychology, education, and clinical practice.

Acknowledgments

This research was supported by a grant from the Natural Sciences and Engineering Research Council (RGPGP-2014-00058 to Marie Arsalidou and Juan Pascual-Leone) and the Russian Science Foundation (17-18-01047 to Marie Arsalidou).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jecp.2022.105462.

References

Agostino, A., Johnson, J., & Pascual-Leone, J. (2010). Executive functions underlying multiplicative reasoning: Problem type matters. *Journal of Experimental Child Psychology, 105*, 286–305.

Arsalidou, M., & Im-Bolter, N. (2017). Why parametric measures are critical for understanding typical and atypical cognitive development. *Brain Imaging and Behavior, 11*, 1214–1224.

Arsalidou, M., Pascual-Leone, J., & Johnson, J. (2010). Misleading cues improve developmental assessment of working memory capacity: The color matching tasks. *Cognitive Development, 25*, 262–277.

Arsalidou, M., Pascual-Leone, J., Johnson, J., Morris, D., & Taylor, M. J. (2013). A balancing act of the brain: Activations and deactivations driven by cognitive load. *Brain and Behavior, 3*, 273–285.

Bilker, W. B., Wierzbicki, M. R., Brensinger, C. M., Gur, R. E., & Gur, R. C. (2014). Development of abbreviated eight-item form of the Penn Verbal Reasoning Test. *Assessment, 21*, 669–678.

Bisagno, E., & Morra, S. (2018). How do we learn to "kill" in volleyball? The role of working memory capacity and expertise in volleyball motor learning. *Journal of Experimental Child Psychology, 167*, 128–145.

Calamia, M., Markon, K., Denburg, N. L., & Tranel, D. (2011). Developing a short form of Benton's Judgment of Line Orientation Test: An item response theory approach. *The Clinical Neuropsychologist, 25*, 670–684.

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31*, 141–154.

Case, R. (1992). The role of the frontal lobes in the regulation of cognitive development. *Brain and Cognition, 20*, 51–73.

Demetriou, A., & Spanoudis, G. (2018). *Growing minds: A developmental theory of intelligence, brain, and education*. New York: Routledge.

Duff, S. J., & Hampson, E. (2001). A sex difference on a novel spatial working memory task in humans. *Brain and Cognition, 47*, 470–493.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*, 1040–1048.

Engle, R. W. (2001). What is working-memory capacity? In H. L. Roediger III & J. S. Nairne (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 297–314). Washington, DC: American Psychological Association.

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross (Ed.), *The psychology of learning and motivation* (pp. 145–199). New York: Elsevier.

Felez-Nobrega, M., Foster, J. L., Puig-Ribera, A., Draheim, C., & Hillman, C. H. (2018). Measuring working memory in the Spanish population: Validation of a multiple shortened complex span task. *Psychological Assessment, 30*, 274–279.

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition, 43*, 226–236.

Geiger, J. F., & Litwiller, R. M. (2005). Spatial working memory and gender differences in science. *Journal of Instructional Psychology, 32*, 49–57.

Gerlach, C., & Poirel, N. (2018). Navon's classical paradigm concerning local and global processing relates systematically to visual object classification performance. *Scientific Reports, 8*, 324.

Gomar, J. J., Harvey, P. D., Bobes-Bascaran, M. T., Davies, P., & Goldberg, T. E. (2011). Development and cross-validation of the UPSA short form for the performance-based functional assessment of patients with mild cognitive impairment and Alzheimer disease. *American Journal of Geriatric Psychiatry, 19*, 915–922.

Goss, P., & Sonnemenn, J. (2017). *Engaging students: Creating classrooms that improve learning*. Carlton, Australia: Grattan Institute.

Guevara, M. A. (2018). *The neural underpinnings of mental attention capacity*. York University. Master's thesis.

Halford, G. S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: A new hypothesis. *Trends in Cognitive Sciences, 11*, 236–242.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences, 21*, 803–831.

Harvey, P. D., Keefe, R. S., Patterson, T. L., Heaton, R. K., & Bowie, C. R. (2009). Abbreviated neuropsychological assessment in schizophrenia: Prediction of different aspects of outcome. *Journal of Clinical Experimental Neuropsychology, 31*, 462–471.

Horton, D. K., Hynan, L. S., Lacritz, L. H., Rossetti, H. C., Weiner, M. F., & Cullum, C. M. (2015). An abbreviated Montreal cognitive assessment (MoCA) for dementia screening. *The Clinical Neuropsychologist, 29*, 413–425.

Howard, S. J., Johnson, J., & Pascual-Leone, J. (2013). Measurement of mental attention: Assessing a cognitive component underlying performance on standardized intelligence tests. *Psychological Test and Assessment Modeling, 55*, 250–273.

Im-Bolter, N., Agostino, A., & Owens-Jaffray, K. (2016). Theory of mind in middle childhood and early adolescence: Different from before? *Journal of Experimental Child Psychology, 149*, 98–115.

Im-Bolter, N., Johnson, J., & Pascual-Leone, J. (2006). Processing limitations in children with specific language impairment: The role of executive function. *Child Development, 77*, 1822–1841.

Johnson, J., Im-Bolter, N., & Pascual-Leone, J. (2003). Development of mental attention in gifted and mainstream children: The role of mental capacity, inhibition, and speed of processing. *Child Development, 74*, 1594–1614.

Kleka, P., & Paluchowski, W. J. (2017). Shortening of psychological tests—Assumptions, methods and doubts. *Polish Psychological Bulletin, 48*, 516–522.

Mann, J. (2018). Is school working for teenage boys? Outdoor learning and real-life skills could be the keys to re-engagement. *Curriculum Perspectives, 38*, 169–174.

Mara, C. A., & Cribbie, R. A. (2018). Equivalence of population variances: Synchronizing the objective and analysis. *Journal of Experimental Education, 86*, 442–457.

Martin, A. J. (2004). School motivation of boys and girls: Differences of degree, differences of kind, or both? *Australian Journal of Psychology, 56*, 133–146.

Martin, A. J. (2007). Examining a multidimensional model of student motivation and engagement using a construct validation approach. *British Journal of Educational Psychology, 77*, 413–440.

Martin, A. J. (2009). Motivation and engagement across the academic life span: A developmental construct validity study of elementary school, high school, and university/college students. *Educational and Psychological Measurement, 69*, 794–824.

Morra, S. (1994). Issues in working memory measurement: Testing for M capacity. *International Journal of Behavioral Development, 17*, 143–159.

Morra, S., Camba, R., Calvini, G., & Bracco, F. (2013). Do Italians do it better? M-capacity measurement and cross-linguistic differences in the Direction Following Task (DFT). *Rivista di Psicolinguistica Applicata, 13*, 9–24.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology, 9*, 353–383.

Onwumere, O., & Reid, N. (2008). Field dependency and performance in mathematics. *European Journal of Educational Research, 3*, 43–57.

Panesi, S., & Morra, S. (2020). Executive functions and mental attentional capacity in preschoolers. *Journal of Cognition and Development, 21*, 72–91.

Pascual-Leone, J. (1970). A mathematical model for the transition rule in Piaget's developmental stages. *Acta Psychologica, 32*, 301–345.

Pascual-Leone, J. (1987). Organismic processes for neo-Piagetian theories: A dialectical causal account of cognitive development. *International Journal of Psychology, 22*, 531–570.

Pascual-Leone, J. (1995). Learning and development as dialectical factors in cognitive growth. *Human Development, 38*, 338–348.

Pascual-Leone, J. (2012). Piaget as a pioneer of dialectical constructivism: Seeking dynamic processes for human science. In E. Martí & C. Rodriguez (Eds.), *After Piaget* (pp. 15–41). New Brunswick, NJ: Transaction.

Pascual-Leone, J. (2014). Dialectics. In T. Teo (Ed.), *Encyclopedia of critical psychology* (pp. 421–428). New York: Springer Reference.

Pascual-Leone, J., & Baillargeon, R. (1994). Developmental measure of mental attention. *International Journal of Behavioral Development, 17*, 161–200.

Pascual-Leone, J., & Ijaz, H. (1989). Mental capacity testing as a form of intellectual developmental assessment. In R. J. Samuda, S. L. Kong, J. Pascual-Leone, J. Cummings, & J. Lewis (Eds.), *Assessment and placement of minority students: A review for educators* (pp. 143–171). Toronto: Hogrefe.

Pascual-Leone, J., & Johnson, J. (1991). The psychological unit and its role in task analysis. A reinterpretation of object permanence. In M. Chandler & M. Chapman (Eds.), *Criteria for competence: Controversies in the assessment of children's abilities* (pp. 153–187). Hillsdale, NJ: Lawrence Erlbaum.

Pascual-Leone, J., & Johnson, J. (1999). A dialectical constructivist view of representation: Role of mental attention, executives, and symbols. In I. E. Sigel (Ed.), *Development of mental representation: Theories and applications* (pp. 169–200). Mahwah, NJ: Lawrence Erlbaum.

Pascual-Leone, J., & Johnson, J. (2005). A dialectical constructivist view of developmental intelligence. In O. Wilhelm & R. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 177–201). Thousand Oaks, CA: Sage.

Pascual-Leone, J., & Johnson, J. (2011). A developmental theory of mental attention: Its applications to measurement and task analysis. In P. Barrouillet & V. Gaillard (Eds.), *Cognitive development and working memory: A dialogue between neo-Piagetian and cognitive approaches* (pp. 13–46). New York: Psychology Press.

Pascual-Leone, J., & Johnson, J. M. (2021). *The working mind: Meaning and mental attention in human development*. Cambridge, MA: MIT Press.

Pascual-Leone, J., Johnson, J., Baskind, S., Dworsky, S., & Severtson, E. (2000). Culture-fair assessment and the processes of mental attention. In I. A. Kozulin & Y. Rand (Eds.), *Experience of mediated learning: An impact of Feuerstein's theory in education and psychology* (pp. 191–214). New York: Pergamon.

Powell, T. L., Arsalidou, M., Vogan, V. M., & Taylor, M. J. (2014). Letter and colour matching tasks: Parametric measures of developmental working memory capacity. *Child Development Research, 2014* 961781.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Robinson, A. (2016). May 14). *Package "equivalence"*. Comprehensive R Archive. *Network.* https://cran.r-project.org/web/packages/equivalence/equivalence.pdf.

Sherrington, C. S. (1940). *Man on his nature*. London: Cambridge University Press.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*, 102–111.

Vanderploeg, R. D., LaLone, L. V., Greblo, P., & Schinka, J. A. (1997). Odd–even short forms of the Judgment of Line Orientation Test. *Applied Neuropsychology, 4*, 244–246.

Vogan, V. M., Morgan, B. R., Lee, W., Powell, T. L., Smith, M. L., & Taylor, M. J. (2014). The neural correlates of visuo-spatial working memory in children with autism spectrum disorder: Effects of cognitive load. *Journal of Neurodevelopmental Disorders, 6*, 19.

Vogan, V. M., Morgan, B. R., Powell, T. L., Smith, M. L., & Taylor, M. J. (2016). The neurodevelopmental differences of increasing verbal working memory demand in children and adults. *Developmental Cognitive Neuroscience, 17*, 19–27.

Vogan, V. M., Morgan, B. R., Smith, M. L., & Taylor, M. J. (2019). Functional changes during visuo-spatial working memory in autism spectrum disorder: 2-year longitudinal functional magnetic resonance imaging study. *Autism, 23*, 639–652.

Weaver-Hightower, M. (2003). The "boy turn" in research on gender and education. *Review of Educational Research, 73*, 471–498.

Wellek, S. (2003). *Testing statistical hypotheses of equivalence*. Boca Raton, FL: Chapman & Hall/CRC.

Woodard, J. L., Benedict, R. H., Roberts, V. J., Goldstein, F. C., Kinner, K. M., Capruso, D. X., & Clark, A. N. (1996). Short-form alternatives to the Judgment of Line Orientation Test. *Journal of Clinical and Experimental Neuropsychology, 18*, 898–904.

Wright, C. V., Beattie, S. G., Galper, D. I., Church, A. S., Bufka, L. F., Brabender, V. M., & Smith, B. L. (2017). Assessment practices of professional psychologists: Results of a national survey. *Professional Psychology: Research and Practice, 48*, 73–78.