The Head-Toes-Knees-Shoulders Revised (HTKS-R): Development and psychometric properties of a revision to reduce floor effects.

Christopher R. Gonzales, Ph.D. [a]

Ryan Bowles, Ph.D. [b]

G. John Geldhof, Ph.D. [c]

Claire Cameron, Ph.D. [d]

Alexis Tracy, M.S. [c]

Megan M. McClelland, Ph.D. [c]

In Press (Accepted March 25, 2021): Early Childhood Research Quarterly

[a] Center for Mind and Brain, University of California, Davis. [b] Human Development and Family Studies, Michigan State University. [c] Human Development and Family Sciences, Oregon State University. [d] Learning & Instruction, University at Buffalo (SUNY).

The research reported here was supported by the U.S. Department of Education Institute for Education Sciences grants # R305A150192 (PI: McClelland) to Oregon State University. The content is the responsibility of the authors and does not necessarily represent the official views of the Institute of Education Sciences, or the U.S. Department of Education. Correspondence concerning this article should be addressed to Christopher Gonzales (chrgonzales@ucdavis.edu) or Megan McClelland (megan.mcclelland@oregonstate.edu).

Highlights

- Self-regulation measures may not capture variability in young children.
- The Head-Toes-Knees-Shoulders-Revised adds a new section to reduce task demands.
- The revised task is reliable and valid in young children.
- The revised task is more related to other self-regulation measures than prior tasks.

Abstract

Measures of self-regulation may not capture adequate variability in children with low levels of self-regulation. This can limit a measure's ability to accurately demonstrate relations with other variables. The present study addressed this issue with a revised version of the Head-Toes-Knees-Shoulders task (HTKS-R), which includes a new downward extension with reduced task demands. Preschool children (*N* = 180; 53% Female) enrolled in Head Start programs were tested with the HTKS-R and other self-regulation tasks at four time points between fall of preschool and spring of kindergarten. Results demonstrated a substantial increase in variability in children's performance on the HTKS-R compared to prior versions of the task during the fall of preschool, and significant increases in the relation between the HTKS-R and other measures of self-regulation at all four time points. Additionally, item factor analyses revealed that the new downward extension captured the same underlying construct as the rest of the measure and that a one factor solution was optimal. Together, these findings illustrate that the HTKS-R improves upon prior versions of the task increasing its utility for researchers and practitioners.

The Head-Toes-Knees-Shoulders Revised (HTKS-R): Development and psychometric properties of a revision to reduce floor effects.

Self-regulation is associated with a variety of social and academic outcomes across the life span (Duckworth, Tsukayama, & May, 2010; McClelland, Acock, Piccinin, Rhea, & Stallings, 2013; McClelland et al., 2007; Moffitt et al., 2011; Cameron Ponitz, McClelland, Matthews, & Morrison, 2009). Although disagreement exists, self-regulation can be defined as a multidimensional construct that incorporates emotion, cognition, and behavior (McClelland, Cameron Ponitz, Messersmith, & Tominey, 2010). The measurement of young children's self-regulation is also greatly varied, ranging from parent and teacher reports of typical behaviors, laboratory- or classroom-based observations, and direct assessments (Blair, Zelazo, & Greenberg, 2016). Each type of measurement exhibits its own strengths and weaknesses, however, direct assessments such as the Head-Toes-Knees-Shoulders task (HTKS), are consistent and strong predictors of child outcomes in preschool and early elementary school (Connor et al., 2016; Fuhs, Nesbitt, Farran, & Dong, 2014; Gestsdottir et al., 2014).

The HTKS is a measure of behavioral self-regulation that captures aspects of executive function (EF; inhibitory control, cognitive flexibility, working memory) manifested in behaviors similar to what children need to do in the classroom such as paying attention, remembering instructions, or stopping one action and doing another. Research on the HTKS suggests it is a reliable and valid measure of behavioral self-regulation for children between four and eight years of age (e.g., McClelland et al., 2014; von Suchodoletz et al., 2013; Wanless et al., 2011), and its use has been extended to include older adults (Cerino et al., 2018). In the task, participants are taught pairs of behavioral commands (e.g., "when I say touch your head, you touch your toes" and vice-versa) which increase in number and complexity as participants progress through the task (i.e., remembering two pairs of commands and then switching the pairs around). Performance on the HTKS is moderately to strongly correlated with other measures of behavioral self-regulation and EF and is consistently a strong indicator in latent variable models (Allan & Lonigan, 2011; Schmitt, et al 2017). However, like many direct measures of behavioral self-regulation, the HTKS demonstrates difficulty in capturing adequate levels of variability in younger children (i.e., a disproportionate floor effect; Cameron Ponitz et al., 2008; McClelland et al., 2014; Schmitt, Geldhof, Purpura, Duncan, & McClelland, 2017) and children from disadvantaged backgrounds (Caughy, Mills, Owen, & Hurst, 2013; Pears, Kim, Healey, Yoerger, & Fisher, 2015). In the current study, we evaluated the psychometric properties and construct validity of a revised version of the HTKS task (the Head-Toes-Knees-Shoulders – Revised; HTKS-R) that adds a downward extension to better capture variability among children with low levels of behavioral self-regulation.

Definitions of Self-Regulation

The term “self-regulation” can be used to describe a range of constructs and skills, and considerable debate remains about its precise definition, constituent components, and relation to similar constructs such as effortful control and EF (Allan & Lonigan, 2011; Eisenberg, Valiente, & Eggum, 2010; Garon, Bryson, & Smith, 2008). Definitions of self-regulation often include both top-down (e.g., EF) and bottom-up regulation of thoughts, feelings, and behavior (Blair & Raver, 2012; Zelazo & Cunningham, 2007); however, differences in these conceptualizations reflect the wide variety of fields that examine self-regulation and the developmental and contextual frameworks in which self-regulation has been considered (e.g., Blair, 2010). Whereas researchers often distinguish among the self-regulation of emotions, cognitions, and behaviors (Calkins, 2007; McClelland, Cameron, Wanless, & Murray, 2007; McClelland et al., 2015), in the current study, we focus on children's behavioral aspects of self-regulation and their relation to underlying EF processes measured with assessments that tap EF components (inhibitory control, cognitive flexibility, working memory). Specifically, we examine children's ability to apply their EF to behavioral responses that are especially relevant in school settings, such as paying attention to instructions when completing behavioral and academic tasks (Cameron Ponitz et al., 2008; Cameron Ponitz, McClelland, Mathews, & Morrison, 2009; Connor et al., 2010; McClelland, Cameron, Wanless, & Murray, 2007; Morrison, Cameron Ponitz, & McClelland, 2010).

Executive functioning is characterized as a multifaceted construct consisting of several related but separable cognitive skills: inhibitory control, attentional or cognitive flexibility, and working memory (Diamond, 2013; Garon et al., 2008; Miyake & Friedman, 2012). Inhibitory control helps children suppress impulsive, or prepotent actions in favor of more adaptive ones (Carlson & Moses, 2001; Dowsett & Livesey, 2000; Rennie, Bull, & Diamond, 2004). Attentional or cognitive flexibility allows children to shift focus back and forth between different task demands and pay attention to new details (Rueda, Posner, & Rothbart, 2005) in ways that facilitate desired behaviors. Working memory allows children to remember, update, and follow directions and allows the planning of solutions to problems (Gathercole & Pickering, 2000; Unsworth & Engle, 2007). Within an integrative self-regulatory framework, these skills overlap with one another, and children utilize them all to better execute the appropriate behavior for a situation (Garon et al., 2008). Thus, it is the co-development of these skills that grants children an increasing capacity to plan, organize, and problem-solve in a variety of contexts so they may better manage their emotions, thoughts, and behavior (Best & Miller, 2010).

The Development of EF and Self-Regulation

There is also age-related variability in how self-regulation is conceptualized and how the underlying EF processes develop (Lee, Bull, & Ho, 2013; Miyake & Friedman, 2012). In studies utilizing a latent variable approach, EF can be best characterized as an integrated, more domain-general construct in early childhood (Hughes, Ensor, Wilson, & Graham, 2010; Wiebe et al., 2011). During the early school-age years, a more complex model that separates inhibitory control from attention shifting and working memory is often more appropriate (Lee, Bull & Ho, 2013), whereas with older children, adolescents and adults, an even more complex model that separates children's inhibitory control, attention-shifting and working memory is often the preferred solution (Lehto et al, 2003; Miyake et al, 2000).

Aspects of EF are believed to be present early in life (Best & Miller, 2010) but develop at different rates across childhood. Children's inhibitory control is known to rapidly develop within the first few years of life (Posner & Rothbart, 1998), whereas children's attentional flexibility and working memory show greater development later in childhood and become increasingly differentiated from one another after age 5 or 6 and into adolescence (Lee et al., 2013). Relations to complex EF tasks that tap multiple components also appear to change across development, with children's performance on complex EF tasks predicted most strongly by measures of inhibitory control early in childhood and by attention shifting and working memory later in childhood and in adolescence (Senn, Espy, & Kaufmann, 2004). Thus, researchers need a better understanding of how measures of self-regulation that require multiple EF components function across development. Particularly important is capturing variability for children with lower skill levels: EF-based measures of self-regulation that emphasize mainly attention shifting or working memory may not be appropriate for younger or lower-skilled children and may result in floor effects (Best & Miller, 2010). To expand the generalizability of complex EF measures during the early childhood period requires measures that include components deliberately designed to assess inhibitory control.

Measuring behavioral self-regulation in early childhood

A considerable source of confusion arises from the ways that the constructs of self-regulation are measured: from direct assessments tapping multiple cognitive skills, to observation systems of overt behavior in both structured and unstructured scenarios, to parent and teacher reports of typical behaviors (Best & Miller, 2010; Morrison & Grammer, 2016). Most direct assessments of preschoolers' self-regulation were developed in laboratory settings and later adapted for use in educational contexts. Carlson (2005) described EF measures available for children aged 2 to 6 years and found that many measures exhibited a binary (pass/fail) distribution. Since then, researchers have developed new measures that capture multiple components of children's executive functioning to help address these issues; however, a number of practical and psychometric problems remain. These issues include the need for specialized materials, training and/or substantial time to administer (Hughes, 1998) or responses which require fine motor actions such as a key press on a laptop or touch screen device (Zelazo et al., 2013).

There are newer measures besides the HTKS that have addressed some of these issues: the Willoughby/Blair battery, (Willoughby, Blair, Wirth, & Greenberg, 2010, 2012; Willoughby, Wirth, & Blair, 2012), The Minnesota Executive Functioning Scale (MEFS; Carlson & Zelazo, 2014), and the NIH Toolbox Cognition Battery (Zelazo, et al., 2013). As with any measure, each of these assessments has strengths and weaknesses. For example, the EF battery (from Willoughby/Blair) comprehensively assesses cognitive flexibility, working memory, and inhibitory control in children aged 3-5 years. The main drawback of the EF Battery is the administration time (up to 45 minutes) and the fact that some components show floor effects (Willoughby, et al., 2010; Willoughby, Blair, et al., 2012; Willoughby, Wirth, et. al., 2012). Similarly, the MEFS (Carlson & Zelazo, 2014) and the NIH Toolbox Cognition Battery (Zelazo, et al., 2013) includes measures of children's executive functioning in computer-adaptive formats which allows for shorter administration times, but administration requires a touch-screen device using a dedicated app with either an annual subscription fee or per-participant costs.

The HTKS is another widely used and validated measure of behavioral self-regulation which addresses many of these shortcomings. The task is short (five to seven minutes to complete), requires no special materials to administer, and has demonstrated good inter-rater reliability when collected by either teachers or trained experimenters (McClelland & Cameron, 2012; Cameron Ponitz et al., 2009). Children's performance on the HTKS is also consistently correlated (e.g., ranging from .20 - .48) with teacher and caregiver ratings of children's self-regulation in preschool and kindergarten years (McClelland et al., 2007; Cameron Ponitz et al., 2009; Schmitt, Pratt & McClelland, 2014), which is similar or higher than other direct assessments of children's EF (Allan et al., 2014). Scores on different versions of the HTKS also have had strong levels of predictive validity both concurrently and longitudinally for several social and academic achievement outcomes in early childhood (McClelland, Cameron, Connor, et al., 2007), in the transition to more formal education settings (Duncan et al., 2007), and beyond (McClelland, et al., 2013). In addition, the HTKS has demonstrated strong construct and predictive validity in cultures beyond the U.S., including samples in Asia (e.g., Taiwan, South Korea, and China; Wanless et al., 2011, 2013), and Europe (e.g., Iceland, Germany, France and Norway; Gestsdottir et al., 2014; Lenes, Gonzales, Størksen & McClelland, 2020; von Suchodoletz et al., 2013; Størksen, Ellingsen, Wanless, & McClelland, 2015).

Despite its relatively simple format, the HTKS is a complex self-regulation task that taps multiple aspects of EF, as demonstrated by relations to measures of working memory, inhibitory control, and cognitive flexibility (Lan, Legare, Cameron Ponitz, Li, & Morrison, 2011; McClelland et al., 2007, 2014). Children utilize several cognitive skills during the task: (1) paying attention to the instructions, (2) remembering the rules of the game while listening to the experimenter give commands, (3) inhibiting the more natural response to the experimenters instruction in favor of the correct, unnatural one, (4) flexibly shifting between old and new rules to commands when given, and (5) regulating their behavior in accordance with all of the above. It is unclear if self-regulation as measured by the HTKS should be considered as reflecting a single dimension or multiple dimensions. Thus, one goal of the present study was to examine the psychometric properties and factor structure of a revised version of the HTKS (e., HTKS-R).

Current Limitations of the HTKS

The HTKS presents two potential sources of difficulty for children with low levels of self-regulation that may account for floor effects in younger children and children from disadvantaged backgrounds. The first is the gross-motor demands of the task in addition to the cognitive demands of attending to and processing instructions. Cognitive and motor performance are interrelated, and evidence increasingly points to motor and cognitive skills drawing from a similar set of cognitive resources (Diamond, 2000; Soska, Adolph, & Johnson, 2010). Thus, a key contributor to performance on many EF tasks may include the ability and necessary resources to properly plan out and execute a correct gross motor response when providing an answer. Considering the mutual resources called upon for planning gross motor behaviors and self-regulation (Boudreau & Bushnell, 2000), the simultaneous demands of the HTKS task may be especially challenging for children with lower self-regulation skills.

The second potential source of difficulty in the HTKS is the social expectation set up between the child and experimenter during the task. Many cognitive tasks require a context in which an adult tells a child what to do (Morrow & Richards, 1996). Children are expected to follow adults' instructions in most early childhood settings, but some self-regulation tasks ask children to deliberately do the opposite of an adult's instruction. Some children might fully understand that disobeying one instruction in favor of another is part of the task and is an acceptable behavior, but other children may be less able to override the prepotent response to do what an adult says at face value. In short, the HTKS may be more challenging because it asks children to disobey an adult.

The HTKS-R was designed to address these two specific issues and to capture additional variability in children with low self-regulation by adding a new downward extension at the beginning of the measure. The new extension contains items with features that reduce the motor and social demands of the task. To reduce the gross motor demands, the new downward extension consists of simple one-word verbal responses with no gross motor component, which was expected to help children focus on the self-regulation aspects of the task. To reduce the social demands, the instructions no longer ask children to disobey what the experimenter tells them to do. In addition, the inclusion of additional reminder prompts scaffolds children's understanding of the task. We expected these modifications would make the downward extension in the HTKS-R easier for children with lower self-regulation skills or younger children in preschool.

Evaluating the HTKS-R

The HTKS has consistently demonstrated strong internal reliability across numerous studies and a diversity of populations (e.g., McClelland et al., 2014; Wanless et al., 2011). The goal of the downward extension in the HTKS-R was to increase the overall sensitivity to variability in children's scores at the lower end of the measure. However, a potential problem with the addition of the downward extension to the HTKS-R is whether the new items in the measure capture additional variability compared to the rest of the measure or if new items measure a different construct.

In general, an important aspect of scale construction is evaluating how individual items within a measure relate to one another (Kline, 1994). By changing the response format and other surface-level features, it is possible that variance in the new items in the downward extension will be unrelated to the variance in children's behavioral self-regulation as measured by the existing items in the HTKS. Thus, this study evaluated how the new downward extension in the HTKS-R relates to the rest of the measure and examined if a unidimensional structure is appropriate (e.g., do the new items tap the same construct as the rest of the measure). A standard way to evaluate the internal structure of a measure is through Item Factor Analysis (IFA), which is a type of exploratory factor analysis used in scale construction with categorical or dichotomously scored items. An IFA provides both a measure of communalities between scale items through standardized factor loadings as well as measures of model fit when evaluating the dimensionality of the overall scale (Wirth & Edwards, 2007). Thus, evaluating changes in relative model fit and the differences the pattern of item factor loadings across different solutions can address how well the new downward extension relates to the rest of the measure.

The Current Study

The goal of the current study was to evaluate the scale and item-level psychometric properties as well as the construct validity of the HTKS-R to help establish the added utility of the measure with children demonstrating low behavioral self-regulation. To accomplish this goal, we sought to answer three research questions. First, we evaluated whether the new downward extension in the HTKS-R captured additional variability in children scoring low on self-regulation by examining the distributional properties of the HTKS-R at the scale level. Second, we examined the psychometric properties of the HTKS-R by using IFA to examine how the new downward extension of the HTKS-R related to the rest of the measure at the individual item level. Third, we evaluated the added value of the HTKS-R compared to the HTKS by examining concurrent and longitudinal relations between the HTKS-R and HTKS in preschool and other measures requiring self-regulation in preschool and kindergarten.

Method

Participants

A total of 180 preschool children (53% female) participated as part of an ongoing study evaluating the measurement of self-regulation in children from low-income families during the transition from preschool to kindergarten. Participants were recruited from Head Start centers in the Pacific Northwest of the United States. To participate in the study, children's families had to meet the federal poverty guidelines for low to moderate family income to be enrolled in a Head Start program and children also had to be eligible to enter kindergarten the following year. Children were assessed in the fall and spring of each year. In the fall of preschool, children were nested within 26 classrooms with an average of 6.73 participating children per classroom (range: 3-12 children). In the fall of kindergarten, children disbursed into 68 classrooms with an average of 1.96 participating children per classroom (range: 1-7 children). Information on children's home language received from the consent form identified children as ELLs for whom Spanish was the primary language spoken at home (26%, 21%, 19%, and 18% at each time point respectively). Spanish-speaking research assistants administered the Pre-Language Assessment

System (preLAS; Duncan & De Avila, 1985-1987) to determine whether a child should receive direct assessments in English or Spanish at each timepoint (16%, 8%, 6% and 6% respectively). Parents reported their child's race as White (51%), Latinx (25%), or Middle Eastern (1%). Only a small percentage of families indicated multiple ethnicities. For example, 10% indicated White and Latinx, 5% indicated White and African American, 4% indicated White and Asian and the remaining 4% were Middle-Eastern and African American or "other". Parents also reported the total number of years of education they completed (M = 11.67 years).

Procedure

Children were assessed on all direct assessments by trained and certified research assistants up to 4 times: in the fall and spring of the preschool and kindergarten years. The order of measures was counterbalanced. Missingness associated with attrition is described below. Assessments were given in short, 10-15-minute sessions and were completed within three classroom visits. Children were identified as Spanish-speaking English Language Learners (ELLs) via the Simon Says and Art Show subtests of the pre-Language Assessment Screener (preLAS), which has demonstrated strong reliability and validity in Spanish-speaking preschool aged children (Rainelli et. al., 2017). Those children identified as Spanish-Speaking, were given the preLAS and assessed by trained Spanish-speaking research assistants at each time point.

Measures

Direct measures of self-regulation.

Children's behavioral self-regulation was measured using a battery of behavioral self-regulation tasks: the HTKS-R, the Day-Night Stroop Task, the Dimensional Change Card Sort Task, and the Woodcock-Johnson (III) Auditory Working Memory Test.

The HTKS-R assesses aspects of cognitive flexibility, working memory, and inhibitory control (McClelland et al., 2014). This task is an updated version of the 3-part HTKS (McClelland et al., 2014) that includes a new downward extension at the beginning of the task for use with children ages 4-8. In the new downward extension, Part 0 (Opposites), there are four practice items and seven test items where children are asked to *say* the opposite of what is instructed (e.g. “if I say toes, you say head”). The rest of the HTKS-R follows the same protocol as the original HTKS with three parts that include practice and test items and ask children to *do* the opposite of what they are told (e.g. touch their toes when asked to touch their head). In Part 1, children are asked to *do* the opposite of what is instructed for a single pair of commands (e.g., “if I say touch your toes, touch your head”). In Part 2, children are given an additional pair of commands to remember (e.g., “if I say touch your knees, touch your shoulders”), and in Part 3, the pairs of commands are switched (e.g., “toes go with shoulders, and knees go with head”). There was a total of 59 items (22 practice items and 37 test items) across the four parts (Parts 0-3). On each item, incorrect responses were scored as 0, self-corrected response were scored as 1, and correct responses were scored as 2. Total scores range from 0-118. The HTKS has demonstrated strong internal consistency in diverse samples around the world (e.g., McClelland et al., 2014; Wanless et al., 2011). In the current sample, the HTKS-R demonstrated strong internal consistency at each time point (Cronbach’s alpha = .97, .97, .97 and 96 respectively). Researchers and practitioners who are interested in implementing the HTKS-R can request access to the measure at https://health.oregonstate.edu/labs/kreadiness/measure.

Children were also assessed using the Day-Night Stroop task (Gerstadt, Hong, & Diamond, 1994), which is designed to primarily tap inhibitory control, and the border-version of the Dimensional Change Card Sort Task (Zelazo, 2006), which is designed to primarily tap cognitive flexibility. In the Day-Night Stroop task, Children are presented with 16 cards with pictures of a sun or moon and asked to say the opposite (e.g., “day” for a moon and “night” for a sun). The measure has demonstrated strong internal reliability in prior research (McClelland et al., 2014; Rhoades, Greenberg, & Domitrovich, 2009). In the current sample, the Day-Night Stroop task demonstrated strong internal consistency (Cronbach’s alpha = .90, .91, .85 and .84 respectively).

The Dimensional Change Card Sort task is separated into three phases (pre-switch, post-switch and borders) with a stopping rules present during between the second and third part (Zelazo, 2006). Children are presented with cards with different colored shapes on them (red rabbits, blue rabbits, red boats and blue boats) one at a time and asked to sort the cards into two different containers. In the pre-switch phase (six trials), children are first asked to sort the cards by their color. In the post-switch phase (six trials), children are then asked to sort the cards by their shape. If children scored five or more points in the post-switch phase, they moved on to the next section. If not, the task ended. In the border phase (twelve trials), children were asked to sort the cards by their color if it has a border on it or to sort the cards by their shape if it did not have a border on it. In the current sample, the DCCS task demonstrated strong internal consistency (Cronbach’s alpha = .94, .92, .90 and .85 respectively).

The Auditory Working Memory subtest of the Woodcock Johnson Tests of Achievement (WJ-III; Woodcock, McGrew, & Mather, 2001) or The Batería III Woodcock- Muñoz (Muñoz-Sandoval, Woodcock, McGrew, & Mather, 2005) was used to assess children’s ability to remember, cognitively manipulate, and give spoken responses from auditory information. The experimenter read a list of two or more numbers and objects to the child, and the child had to repeat the list back to the experimenter by first saying the numbers in the list in the order that they heard them and then the objects in the list in the order that they heard them. Based on the normed samples, Cronbach's alpha for English-speaking preschool aged children range between .93-.96 and .77 - .79 for Spanish-speaking children (Schrank et al., 2005; Woodcock & Mather, 2000).

Results

We pursued three related research questions in this study. First, we examined the scale-level psychometric properties of the HTKS-R to see if it captured adequate variability in self-regulation with young children from disadvantaged backgrounds. Second, we evaluated the item-level properties of the HTKS-R to see how well the new downward extension related to the rest of the measure. Third, we examined the value that the downward extension added to the HTKS-R in terms of its construct validity by examining how the HTKR relates to other common measures of self-regulation compared to the HTKS.

Analytic Strategy

We addressed each of the research questions using either Stata 15.1 (StataCorp, 2018) or M-Plus 8.4 (Muthén & Muthén, 2017). Because of the nested structure of observations within children and children within classrooms, we first examined whether it was necessary to apply a multilevel framework (Hox, 2010) to our analyses. For each research question, we calculated the Intra-Class Correlation (ICC) of the HTKS-R for children clustered within classrooms at each time point. For research questions one and two, we calculated the ICC of the HTKS-R for observations nested within children. Within each timepoint, the ICCs for children nested within classrooms with the HTKS-R were generally small (.04 - .18) but were within the range where accounting for the nested structure of the data is appropriate (Hox, 2010). The ICCs for observations within children (.30) were also above this range.

For research question one, we included a random effect in the logistic regression models for observations nested within children and when comparing variances in the HTKS and HTKS-R, we utilized residual variances from an intercept only random effects models to account for children being nested within classroom. For research question two, models that included a random effect term for both observations nested within children and children nested within classrooms would not converge when the analysis included all test-items in the measure. The convergence issues were likely due to many children in the analyses with no variability across observations on some items in the measure making it difficult to estimate a random effect at the child level. Thus, models reported below only included clustered-robust standard errors to account for clustering of children within classrooms. However, random-effects models that removed problem items with low variability across observations while also accounting for both the clustering of observations within children and the clustering of children within classrooms largely gave the same result. For research question three, we accounted for clustering of children within classrooms by calculating the effective sample size for each analysis. The effective sample size adjusts the degrees of freedom downward proportionally as a function of the average cluster-size and ICC (Hox, 2010).

Descriptive Statistics, Missing Data, and Attrition. Descriptive statistics and sample sizes for the demographic covariates and the EF measures at each wave are reported in Table 1. Cross-timepoint correlations between the Fall of preschool and the Spring of kindergarten for the HTKS-R ranged from .40 to .68, *ps* < .001. For the Day-Night task, correlations ranged from .14 (p = .123) to .51 (p < .001) . For the DCCS, correlations ranged from .20 (p = .026) to .41 (p < .001), and for the WJ Working Memory task, correlations ranged from .18 (p = .046) to .45 (p < .001).

Missing data on measures was most commonly occurred due to attrition. At the second time point (in the spring of the preschool year), 29 children (16%) did not contribute any data on direct measures. At the third time point (in the fall of kindergarten year), 47 children (26%) did not contribute any data on direct measures, and at the final time point (in the spring of the kindergarten year), 54 children (30%) did not contribute any data on direct measures. We tested whether children's age, gender or ELL-status were significant predictors of attrition at any wave of data collection utilizing a logistic regression model to predict attrition. One covariate was a significant predictor of missingness at Wave 2. Children who were assessed in Spanish at Wave 1 were significantly less likely to participate in Wave 2. No other significant predictors of attrition emerged at other time points.

Missing data not due to attrition was otherwise rare and was mostly due to a child being absent from school during subsequent sessions of data collection, or a child refusing to complete any remaining tasks during a testing session. During the fall of preschool for Wave 1, eight children (4%) who consented and participated in later time points did not contribute any data on direct measures. Within single measures at Wave 1, the most missing data occurred on the Woodcock Johnson auditory working memory task with 7% missing data. At Wave 2, the task with the most missing data was also the Woodcock Johnson Auditory Working Memory task with 15% missing data. At Wave 3, there were no missing data on individual tasks except for one child who did not receive the Dimensional Change Card Sort task. At Wave 4, there were no missing data on individual tasks. No significant predictors emerged for missingness on individual tasks.

To properly model the categorical nature of the outcome variables for research question two, we utilized a weighted least squares estimator with mean and variance correction

(WLSMV) in our models. For research question three, we utilized a full information maximum likelihood (FIML) estimator for continuous outcome variables to account for the small amount of missing data in individual predictors not due to attrition. These approaches to missing data provide a more optimal solution than more traditional (e.g., listwise deletion) missing-data handling techniques under a Missing at Random (MAR) or a Missing at Random with respect to X (MAR-X) assumption (Enders, 2010). Although the WLSMV estimator used in research question two utilizes pairwise deletion, it also provides consistent parameter estimates compared to a FIML estimator under an MAR-X assumption (Asparouhov & Muthén, 2010).

Research Question 1: Variability in the HTKS-R. We first evaluated whether the HTKS-R captured additional variability in children at risk for low self-regulation. As shown in Table 1, children's scores increased over each wave of data collection, and variability was present at each time point. To assess whether the HTKS-R could capture additional variability in children who are both young and from disadvantaged backgrounds compared to the standard HTKS, we further evaluated the variability in children's scores and the proportion of children scoring a zero (i.e., proportion floor effect) on the measure in Fall of preschool in Wave 1. First, we grouped children into four age-groups approximately four months apart from 48 months of age to 60 months of age. Second, to estimate the variability and change in the proportion of floor effects in the HTKS-R, we calculated two different sum scores from children's performance. The first sum score was calculated from the original 48 items from parts 1 to 3 of the measure which were identical to the existing three-part HTKS (McClelland et al., 2014). The second sum score was calculated from all 59 items of the four-part HTKS-R, including the new downward extension.

Table 2 shows the mean scores, skewness and kurtosis, and the proportion floor effect on the HTKS and HTKS-R for all age-groups in Wave 1. A substantial proportion of children scored a zero on the original HTKS items (i.e., sum scores excluding the new downward extension) which ranged from 12.73% to 22.58% with an overall floor effect of 16.57%. This is similar to results in prior studies (Fuhs et al., 2014; McClelland et al., 2014). However, on the HTKS-R (i.e., including the new downward extension), only a few children across any of the age-groups scored a zero (floor effects ranged from 1.96% to 3.64%) with an overall floor effect of 2.96%. Thus, the HTKS-R provided an 82% reduction in the floor effect on the measure compared to the HTKS.

To test if the decrease in the likelihood of children scoring a zero between the HTKS-R and the HTKS was statistically significant, we conducted a logistic regression comparing the odds of not scoring a zero on the HTKS-R with the odds of not scoring a zero on the HTKS at the first time point, while adjusting for repeated measures of observations within children. Analyses revealed that children were significantly more likely to score 0 on the HTKS compared to the HTKS-R, $z = 4.47$, $p < .001$, odds ratio = 7.13, 95% C.I. (3.02 : 16.86). Similarly, an equivalence of variance F-test utilizing residual variances from intercept only random effects models to account for children nested within classrooms revealed that there was significantly more variance in children's scores on the HTKS-R than in the HTKS in the fall of preschool , $F(168, 167) = 2.34$, $p < .001$, and the spring of preschool, $F(136, 135) = 2.02$, $p < .001$, as well as in the fall of kindergarten, $F(132, 131) = 1.78$, $p = .006$, and spring of kindergarten, $F(125, 124) = 1.65$, $p = .028$.

Research Question 2: Psychometric Properties of the HTKS-R. Our second goal was to evaluate the psychometric properties of the HTKS-R by evaluating how children's performance on the downward extension in the HTKS-R related to performance on other sections of the measure at the item-level. To accomplish this goal, we first examined how sum scores across the different parts of the HTKS-R correlated with one another. These correlations and the means and standard deviations for each part of the measure across the four time points are reported in Table 3. In all but one instance across all four time points, each part of the HTKS-R was significantly correlated with all other parts of the measure.

To further examine the item-level properties of the measure, we ran an exploratory IFA. Consistent with prior research supporting multiple factors in children's EF, preliminary analyses in the IFA revealed three factors with eigen values greater than one. Thus, we pursued one, two, and three factor solutions allowing all individual items in the measure to freely load onto each factor.

Because of the large number of indicators in the model, we took several approaches to maximize the statistical power of the analysis and aid in model convergence: (1) we utilized all available data from all four time points for each participant as individual observations in the analysis (N =565) and utilized cluster-robust standard errors to account for clustering of observations within children; (2) to increase the ratio of observations to indicators in the analysis to acceptable levels, we only included the test items in the analysis (k = 37) rather than combining test items and practice items; and (3) because of the small number of self-corrects in the dataset (*M* = 3.56% self-corrects per item, SD = 3.12%), we dichotomized all responses as either correct [1] or incorrect [0] by scoring any self-corrects as a correct response. For the analyses below, some items showed standardized factor loadings greater than one. This can sometimes occur when items are highly correlated (Chen & Deegan, 1978), which would be expected in a measure with very similar test items such as in the HTKS.

One-factor solution. As shown in Table 4, for the one-factor model, all test-items in the HTKS-R loaded significantly onto the single latent factor with factor loadings for test items in the downward extension ranging from .63 to .92, and factor loadings on test-items from Part 1 through 3 ranging from .88 to .99. As shown in the bottom of Table 4, the RMSEA, SRMR, CFI and TLI were all generally within the acceptable to excellent range (West, Taylor, & Wu, 2012), with the CFI and TLI above .99).

Two-factor solution. For the two-factor solution, the earlier and easier parts of the measure loaded more strongly onto the first factor while the later and harder parts of the measure loaded more strongly onto the second factor. More specifically, items from the downward extension all loaded significantly onto the first factor with factor loadings ranging from .75 to 1.05 but not onto the second factor. Test items from Part 1 loaded significantly onto both factors with factor loadings ranging from .43 to .67. Test items for Part 2 loaded significantly onto the second factor with factor loadings ranging from .66 to .85 with smaller but still significant factor loadings on the first factor ranging from .21 to .35. Test items for Part 3 loaded significantly onto the second factor with factor loadings ranging from .87 to 1.06. One test item for Part 3 had a significant negative loading on the first factor. Additionally, the two latent factors were significantly correlated within another, $r = .49$, $p < .05$. Using the DIFFTEST option in Mplus, which provides the appropriate solution with models utilizing a WLSMV estimator (Muthén, du Toit, & Spisic, 1997), we found that the two-factor solution provided a statistically significant improvement over the one factor solution, $\chi(36) = 381.99$, $p < .001$. As shown at the bottom of Table 4, model fit as indexed by the RMSEA, SRMR, CFI and TLI was generally improved in the two-factor solution compared to the one factor solution, although the change in CFI and TLI was negligible and below levels considered as substantial change (Chen, 2007).

Three-factor solution. For the three-factor solution, significant item factor loadings were more dispersed across the three factors. Generally, the downward extension was isolated on its own in the first factor, whereas items from the rest of the HTKS all showed large and significant factor loadings with the second factor. There was no clear pattern of factor loadings on the third factor. More specifically, items from the downward extension all loaded significantly onto the first factor with item loadings ranging from .71 to 1.05. However, one item also loaded significantly onto the second factor with a factor loading of .20. In contrast to the two-factor solution, all items from Part 1 of the measure loaded significantly onto the second factor with factor loadings ranging from .76 to .88. Additionally, all items from Part 1 had significant cross loadings onto the third factor with factor loadings ranging from .29 to .50. For items from Part 2 of the measure, all items loaded significantly onto the second factor with factor loadings ranging from .70 to .85. Moreover, all items had significant cross loadings onto the first factor with factor loadings ranging from .14 to .24. All items from Part 3 had significant factor loadings onto the second factor with factor loadings ranging from .94 to 1.18. Finally, four items from Part 3 showed significant negative cross loadings onto the first factor, and 6 items showed significant negative cross loadings onto the third factor.

All three latent factors were significantly correlated with one another. The first latent factor was significantly correlated with the second latent factor, r = .59, $p < .05$, and with the third latent factor, r = .36, $p < .05$. The second latent factor was also correlated with the third latent factor, r = .18, $p < .05$. As shown on Table 4, similar to the lower-order models, the three-factor solution demonstrated improvements in relative model fit as indexed by the RMSEA, SRMR, CFI and TLI over and above the two-factor solution. We utilized the same DIFFTEST option in MPlus as described above, and tests revealed that the three-factor solution did provide significant improvement in fit over the two-factor solution, $\chi(35) = 241.81$, $p < .001$. However, the pattern of factor loadings made it difficult to interpret the three-factor solution.

Summary. Each of the models demonstrated excellent levels of absolute model fit, and each of the more complex models demonstrated significant improvements in terms of absolute model fit over the prior simpler solution. However, especially with dichotomous items in an IFA, model fit estimates can be sensitive enough to identify under-factored solutions but can struggle to differentiate between over-factored solutions (Clarks & Bowles, 2018). One of the most sensitive estimates of model fit for over-factored solutions is the CFI (Chen, 2007; Clarks & Bowles, 2018). For all three models, the CFI demonstrated excellent levels of relative model fit (>.99), and there were negligible changes in the CFI between each of the three models, suggesting that additional factors beyond a single factor offered no additional explanatory value while offering less parsimony. Thus, we concluded that the one-factor solution represents the optimal solution.

Results for the exploratory IFA within individual time points revealed a similar pattern of results in terms of factor structure and model fit. However, there were some differences in the pattern of factor loadings in the two- and three-factor solutions across timepoints. Due to the underpowered nature of these analyses, these results should be interpreted with caution. Further details of these on these analyses are available in the online supplemental materials.

Research Question 3: Construct Validity of the HTKS-R. Our third goal was to provide an initial evaluation of the construct validity and added value of the HTKS-R by examining the concurrent and longitudinal relations with other measures that tap individual EF components compared to the HTKS. Concurrent and longitudinal correlations between the HTKS and the HTKS-R with other EF measures are reported in Table 5. Results indicated that both the HTKS and the HTKS-R were highly correlated with one another as expected, and significantly correlated with all other measures of EF at each time point. To examine the differences in the correlations between the HTKS and HTKS-R with other measures of EF, we first calculated the effective sample size for each pair of correlations which adjusts the degrees of freedom downward proportionally as a function of the ICC and average cluster size, and then utilized Fisher's r-to-z transformation to test for significant differences between two correlations from dependent samples (Steiger, 1980). As shown on Table 6, analyses revealed the correlations between the HTKS-R and the Day-Night task were significantly larger than the correlations between the HTKS and the Day-Night task within every time point. When examining the relations across timepoints from the Fall of preschool to the Fall of Kindergarten and from the Spring of Preschool to the Spring of Kindergarten, the correlation between the HTKS-R and the Day-Night task were also significantly larger than the correlation between Day-Night task and HTKS.

Additionally, the correlations between the HTKS-R and the DCCS were significantly larger than the HTKS and DCCS within the Fall and Spring of the Kindergarten time points, but not in the preschool time points. However, when examining the relations across timepoints from the Fall of preschool to the Fall of Kindergarten, the correlations between the HTKS-R at the preschool timepoint and the DCCS at the kindergarten timepoint was significantly larger than the correlations with the HTKS. There were no significant differences in the correlations between the HTKS-R and HTKS in relation to the WJ Auditory Working Memory task.

Discussion

The goal of this study was to evaluate the psychometric properties and construct validity of the newly developed HTKS-R in a longitudinal sample of children during the transition to kindergarten. Prior research on the HTKS has demonstrated that, although it is a strong and reliable indicator of children's behavioral self-regulation, the measure does not always adequately capture variability in young children at risk for low self-regulation. The HTKS-R was developed to address this problem by adding a new downward extension to the measure with reduced motor and social demands on children’s performance. Results indicated that the HTKS-R largely accomplishes this goal.

Distributional properties of the HTKS-R at the scale level

At the scale-level, the HTKS-R captured more variability in children’s self-regulation than has been found with the HTKS. With the inclusion of the downward extension in the HTKS-R, the floor-effect of the measure was reduced more than 80% compared to the HTKS with only a few children scoring a zero at the earliest time points of the study. This is important statistically because it increases sensitivity of the measure for studies seeking to correlate the measure with other constructs or skills. There are few behavioral self-regulation measures that are valid for use between ages 4 to 8. Behavioral self-regulation is especially predictive of early literacy and mathematics skills in the preschool period, but patterns can be masked when measures are at floor or ceiling level. If researchers have access to a more sensitive measure of self-regulation, it will be easier to identify unique associations with various aspects of emergent literacy or mathematics. This finding is important practically because, in terms of school readiness assessment, the children for whom practitioners most want to measure self-regulation accurately are those with the lowest skill levels (e.g., Fuhs et al., 2014; Cameron et al., 2012; Purpura et al., 2017) and children from disadvantaged backgrounds (McClelland & Cameron, 2012; Willoughby et al., 2012).

Item-level properties of the HTKS-R

Despite being administered in a different format to reduce motor and social demands, the downward extension in the HTKS-R largely captured the same underlying construct(s) as the existing items in the HTKS. The IFA analysis allowed for the examination of how the new items in the downward extension related to items in the rest of the measure in terms of the simpler, more domain general construct of EF typically found with younger children (Hughes, Ensor, Wilson, & Graham, 2010; Wiebe et al., 2011) as well as the less integrated more complex models of EF that are typically found with older children (Lee, Bull & Ho, 2013; Lehto et al, 2003; Miyake et al, 2000). In terms of interpretability of the different solutions, both the one factor model and the two-factor model in the IFA had patterns of factor loadings that were in line with these conceptual models. In the one factor solution, all items had large factor loadings on the single latent construct. In the two-factor model, the pattern of factor loadings worked out such that items from the new downward extension and the easier items in Parts 1 and 2 loaded strongest onto one latent construct, and the harder items from Part 2 and 3 loaded strongest onto the second latent construct. This pattern of results could be artifactual, reflecting a difficulty factor (McDonald & Ahlawat, 1974), rather than a substantive difference. Alternatively, this pattern of results could support the notion that children's inhibitory control becomes increasingly differentiated from their attention shifting and working memory-ability (Lee, Bull & Ho, 2013),

We chose the simpler one factor model as the more parsimonious, theoretically supported, and practical solution for the purposes of this study. All three solutions tested in the IFA demonstrated good to excellent estimates of model fit. Standard estimates of model fit such as chi-square and RMSEA can readily identify when a more complex model is needed to better explain patterns in the data (Chen, 2007). However, they are less useful when trying to determine if better fitting models are overly complex (Clarks & Bowles, 2018). One of the most sensitive estimates of model fit for over-factored solutions is the CFI (Chen, 2007, Clarks & Bowles, 2018). For all three of the models tested, the CFI consistently showed excellent levels of model fit with negligible changes between each of the models. This supports the argument that the one-factor model represents the most parsimonious and the most practical solution for the purposes of the current study with children ages 4 to 6, although the two-factor model was also acceptable and should be examined in future research.

Construct validity of the HTKS-R in preschool and kindergarten

The final goal of the study was to examine the concurrent and longitudinal relations between the HTKS-R and other measures of tapping individual components of EF by evaluating how performance on the HTKS and HTKS-R in preschool were related to levels of different measures of behavioral self-regulation in kindergarten. Results demonstrated that, similar to previous research on the HTKS (e.g., McClelland et al., 2014), the HTKS-R was significantly related to measures of working memory, inhibitory control, and cognitive flexibility. Although similar in magnitude, relations were also stronger for the HTKS-R compared to the HTKS both within and across timepoints especially for a measure of inhibitory control (on the Day-night task) in fall and spring of preschool and kindergarten and for a measure of cognitive flexibility (the DCCS) in the fall and spring of kindergarten. Because of the high correlation between the HTKS and HTKS-R, additional variance explained by the HTKS-R over and above the HTKS was expected to be small, but still meaningful because it is capturing something that the previous version of the measure could not.

This supports research arguing that the HTKS-R taps aspects of inhibitory control (McClelland et al., 2015) although it also taps aspects of cognitive flexibility and working memory (on the WJ Auditory Working Memory). Although the HTKS-R was significantly related to working memory at all time points, there were no significant differences in the magnitude of the correlations of the HTKS-R and HTKS with working memory. Together, this suggests that compared to the HTKS, the HTKS-R taps all aspects of EF (inhibitory control, cognitive flexibility, and working memory) and is more strongly related to a measure of inhibitory control at all time points and more strongly related to a measure of cognitive flexibility in kindergarten.

Implications

Our results have implications for practice and research. School leaders and teachers need economical, practical, and culture-friendly assessments that can help predict whether children will make a successful transition to formal schooling. Moreover, practitioners need practical tools to identify children who may be at-risk for difficulties as they enter school. There are other measures of self-regulation and EF that are also appropriate for school settings. For example, the MEFS is a measure that has been used in schools although it requires a tablet or computer and there is a per-participant fee (Carlson & Zelazo, 2014). The HTKS-R represents a short (5-7 minutes) and easy-to-administer option that can be used without materials such as computers or tablets and can identify children who may have difficulty with the self-regulatory skills that have been shown to predict academic achievement. Thus, from a practical perspective, the HTKS-R may be a useful tool that captures children’s readiness for the demands of more structured classrooms that are common in elementary school.

In terms of research, the existing HTKS has become a useful research tool around the world. A large number of studies have used the HTKS in many cultures and contexts, and the measure is available in over 27 languages (e.g., Gestsdottir et al., 2014; McClelland et al., 2014; Wanless et al., 2011; von Suchodoletz et al., 2013). The results of the present study supporting the psychometric properties of the new HTKS-R suggest that the HTKS-R improves upon the existing HTKS and can be an effective tool for school readiness assessments and research. However, research using the HTKS-R in other samples and cultures is needed. Developing norms so practitioners can interpret a child's score in relation to scores of other children is also a critical next step. Other measures such as the NIH Toolbox are also important research tools (Zelazo et al., 2013). However, the measures in the NIH Toolbox require a computer or tablet, cost money, and may pose administration barriers for use in school-based settings. Thus, the HTKS-R could be a practical alternative for researchers who need a short and easy to administer measure that does not cost money or require other materials (e.g., computers or tablets).

To further illuminate the conceptual and theoretical questions our study raises, researchers should continue to examine the factor structure of the HTKS-R among samples with a wider developmental span and from diverse backgrounds. Our results with children ages 4 to 6 are consistent with unidimensional solutions found in other studies of behavioral self-regulation and EF measures during the formal school transition period (e.g., Schmitt et al., 2017; Wiebe et al., 2011). The implications of the unidimensionality of children's EF for practitioners are straightforward: supporting young children's behavioral self-regulation can be accomplished through activities that support any of the individual EF components, though it may be easier and most developmentally appropriate to focus on measures that are unidimensional for the youngest children; in other words, by placing greater emphasis on children's inhibitory control.

Limitations and Future Directions

The results of the current study offer several insights into the psychometric properties of the HTKS-R as a complex measure of children's self-regulation, but there are limitations that need to be addressed in future research. First, results are based on a sample of children from similar backgrounds (i.e., there was sample homogeneity as all families qualified for Head Start and were low-income), resulting in a normative sample of children assessed in their primary language. This limitation was intentional as the HTKS-R was designed to address the shortcomings of the HTKS with children who have low levels of self-regulation, including younger children whose self-regulation is still developing as well as an overrepresentation of children from low-income or otherwise disadvantaged families, given that factors such as poverty, racism, and other inequalities make it difficult for children to develop the regulatory skills that are demanded by early school contexts (Blair & Raver, 2012). The HTKS-R was also designed to reduce the social demands of the task while placing minimal additional demands on children's verbal ability. Nevertheless, future research should continue to evaluate properties of children's performance on the HTKS-R in more diverse samples to ensure that measure continues to function as intended, including samples of children with varying verbal abilities and children in different cultural contexts.

Second, the somewhat small sample size may have limited the generalizability of our findings and the ability to detect significant effects especially when trying to predict later developing aspects of children's self-regulation. Future studies with more nuanced analyses should explore how longitudinal differences in early performance on the HTKS-R or differences in the gains on the HTKS-R over time might better relate to these other outcomes. In the current study, we did not have the statistical power to adequately test these hypotheses. When evaluating the dimensionality of the HTKS-R, we had to combine multiple observations of children's performance on the HTKS-R as they developed into a single sample to achieve adequate levels of statistical power for this type of analysis. A limitation of this approach is that it can limit our ability to assess differentiation in children's self-regulation as they matured. Although the single-factor model is appropriate to use in practical scenarios (i.e., it fit well at all ages), there could be more subtle patterns of development that were masked by the wide age range in the sample. For example, for children in the early school-age years, the HTKS-R may be better represented by a two-factor model that differentiates between children's inhibitory control and their attention shifting and working memory-abilities. The under-powered wave-specific models in the online supplemental materials suggest that these subtle nuances that may be more observable in larger samples. Additional data could further inform our understanding of how aspects of children's behavioral self-regulation develop during this age-period.

Finally, whereas the current study provides initial evidence to establish the psychometric properties and construct validity of the HTKS-R as an effective measure of self-regulation in young children, future research should examine the added benefit of the HTKS-R in terms of its predictive validity for important child-outcomes such as their early academic achievement and social-emotional competence. Prior versions of the HTKS have consistently demonstrated strong concurrent and longitudinal associations with such outcomes both as an individual unique predictor as well as an indicator in latent-variable models. It seems likely that the HTKS-R should enhance the ability to detect such relations in future research but needs to be investigated.

Conclusion

This study establishes several psychometric strengths of the HTKS-R measure of behavioral self-regulation among children from low-income families over the transition from preschool to kindergarten. Like the existing HTKS, the HTKS-R is economical and easy to administer. Accessible assessment is a key to supporting the children most marginalized by society. To work well in both research and practical settings, measures also must capture individual differences among young children and those with self-regulation difficulties, as the

HTKS-R does. A single-factor solution combined with significant correlations with individual EF component measures also establishes the HTKS-R as a measure of complex behavioral self-regulation. This is important because previous work shows that complex measures are most predictive of behavioral and academic outcomes. Overall, results support the HTKS-R as a practical and feasible measure of behavioral self-regulation that can be used to support children as they begin their academic careers.

References

Allan, N. P., & Lonigan, C. J. (2011). Examining the Dimensionality of Effortful Control in Preschool Children and Its Relation to Academic and Socioemotional Indicators. *Developmental Psychology*, *47*(4), 905–915. https://doi.org/10.1037/a0023748

Allan, N. P., Hume, L. E., Allan, D. M., Farrington, A. L., & Lonigan, C. J. (2014). Relations between inhibitory control and the development of academic skills in preschool and kindergarten: A meta-analysis. *Developmental Psychology, 50*(10), 2368-2379. https://doi.org/10.1037/a0037493

Asparouhov, T., & Muthen, B. (2010). Weighted least squares estimation with missing data. Retrieved from http://www.statmodel.com/download/GstrucMissingRevision.pdf

Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child Development*, *81*(6), 1641–1660. https://doi.org/10.1111/j.1467-8624.2010.01499.x

Blair, C. (2010). Stress and the development of self-regulation in context. *Child Development Perspectives, 4*(3), 181-188. https://doi.org/10.1111/j.1750-8606.2010.00145.x

Blair, C., & Raver, C. C. (2012). Child development in the context of adversity: experiential canalization of brain and behavior. *The American Psychologist*, *67*(4), 309–318. https://doi.org/10.1037/a0027493

Blair, C., Zelazo, P. D., & Greenberg, M. T. (2016). *Measurement of executive function in early childhood: a special issue of developmental neuropsychology*. Retrieved from https://www.taylorfrancis.com/books/9781135067250

Calkins, S. D. (2007). The emergence of self-regulation: Biological and behavioral control mechanisms supporting toddler competencies. In C. A. Brownell & C. B. Kopp (Eds.), *Socioemotional development in the toddler years: Transitions and transformations* (pp. 261-

284). New York, NY: Guilford.

Cameron, C. E., et al., (2012). Fine motor skills and executive function both contribute to kindergarten achievement. *Child Development, 83*(4), 1229-44. https://doi.org/10.1111/j.1467-8624.2012.01768.x

Cameron Ponitz, C. E., McClelland, M. M., Matthews, J. S., & Morrison, F. J. (2009). A structured observation of behavioral self-regulation and its contribution to kindergarten outcomes. *Developmental Psychology*, *45*(3), 605–619. https://doi.org/10.1037/a0015365

Cameron Ponitz, C.E., McClelland, M.M., Jewkes, A.M., McDonald Connor, C., Farris, C.L., Morrison, F.J. (2008) Touch your toes! Developing a direct measure of behavioral regulation in early childhood. Early Childhood Research Quarterly, 2008. 23(2), 141-158. https://doi.org/10.1016/j.ecresq.2007.01.004

Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology, 28*(2), 595-616. https://doi.org/10.1207/s15326942dn2802_3

Carlson, S. M., & Moses, L. J. (2001). Individual Differences in Inhibitory Control and Children's Theory of Mind. *Child Development, 72*(4), 1032-1053. Retrieved from http://dx.doi.org/10.1111/1467-8624.00333

Carlson, S. M., & Zelazo, P. D. (2014). *Minnesota Executive Function Scale: Test Manual*. Reflection Sciences, Inc. St. Paul, MN.

Caughy, M. O., Mills, B., Owen, M. T., & Hurst, J. R. (2013). Emergent self-regulation skills among very young ethnic minority children: a confirmatory factor model. *Journal of Experimental Child Psychology*, *116*(4), 839–855. https://doi.org/10.1016/j.jecp.2013.07.017

Cerino, E. S., Hooker, K., Stawski, R. S., & McClelland, M. M. (2018). Adapting the Head-Toes-Knees-Shoulders task to older adults: A psychometric evaluation of a measure of processing speed, attention, and inhibitory control. *The Gerontologist.* https://doi.org/10.1093/geront/gny028

Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural Equation Modeling: A Multidisciplinary Journal, 14(3), 464–504. https://doi.org/10.1080/10705510701301834

Chen, F., Deegan, J. (1978). On the Occurrence of Standardized Regression Coefficients Greater Than One. *Educational and Psychological Measurement*, *38*(4), 873–888. https://doi.org/10.1177/001316447803800404

Clark, D. A., & Bowles, R. P. (2018). Model Fit and Item Factor Analysis: Overfactoring, Underfactoring, and a Program to Guide Interpretation. *Multivariate Behavioral Research, 53*(4), 544-558. https://doi.org/10.1080/00273171.2018.1461058

Connor, C. M., Cameron Ponitz, C. E., Phillips, B. M., Travis, Q. M., Glasney, S., & Morrison, F. J. (2010). First graders' literacy and self-regulation gains: The effect of individualizing student instruction. *Journal of School Psychology, 48*, 433-455. https://doi.org/10.1016/j.jsp.2010.06.003

Connor, C. M., Day, S. L., Phillips, B., Sparapani, N., Ingebrand, S. W., McLean, L., . . . Kaschak, M. P. (2016). Reciprocal effects of self-regulation, semantic knowledge, and reading comprehension in early elementary school. *Child Development, 87*(6), 1813-1824. https://doi.org/10.1111/cdev.12570

Diamond, A. (2000). Close interrelation of motor development and cognitive development and of the cerebellum and prefrontal cortex. *Child Development, 71*(1), 44-56.

https://doi.org/10.1111/1467-8624.00117

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Dowsett, S. M., & Livesey, D. J. (2000). The development of inhibitory control in preschool children: Effects of 'executive skills' training. *Developmental Psychobiology, 36*, 161-174. https://doi.org/10.1002/(sici)1098-2302(200003)36:2<161::aid-dev7>3.0.co;2-0

Duckworth, A. L., Tsukayama, E., & May, H. (2010). Establishing Causality Using Longitudinal Hierarchical Linear Modeling: An Illustration Predicting Achievement From Self-Control. *Social Psychological and Personality Science*, *1*(4), 311–317. https://doi.org/10.1177/1948550609359707

Duncan, S. E., & De Avila, E. A. (1985-1987). *PRE-LAS English*. Monterey, CA: CTB/McGraw-Hill.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428-1446. https://doi.org/10.1037/0012-1649.43.6.1428

Eisenberg, N., Valiente, C., & Eggum, N. D. (2010). Self-Regulation and School Readiness. *Early Education and Development*, *21*(5), 681–698. https://doi.org/10.1080/10409289.2010.497451

Enders, C. (2010). *Applied missing data analysis*. New York: Guilford Press.

Fuhs, M. W., Nesbitt, K. T., Farran, D. C., & Dong, N. (2014). Longitudinal associations between executive functioning and academic skills across content areas. *Developmental Psychology, 50*(6), 1698-1709. https://doi.org10.1037/a0036633

Garon, N., Bryson, S. E., & Smith, I. M. (2008). Executive Function in Preschoolers: A Review

Using an Integrative Framework. *Psychological Bulletin*, *134*(1), 31–60. https://doi.org/10.1037/0033-2909.134.1.31

Gathercole, S. E., & Pickering, S. J. (2000). Working memory deficits in children with low achievements in the national curriculum at 7 years of age. *The British Journal of Educational Psychology*, *70 (Pt 2)*, 177–194. https://doi.org/10.1348/000709900158047

Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: performance of children -7 years old on a stroop- like day-night test. *Cognition, 53*(2), 129-153. https://doi.org/10.1016/0010-0277(94)90068-x

Gestsdottir, S., von Suchodoletz, A., Wanless, S. B., Hubert, B., Guimard, P., Birgisdottir, F., … McClelland, M. (2014). Early behavioral self-regulation, academic achievement, and gender: Longitudinal findings From France, Germany, and Iceland. *Applied Developmental Science*, *18*(2), 90–109. https://doi.org/10.1080/10888691.2014.894870

Hox, J. (2010). *Multilevel Analysis* (2nd ed.). https://doi.org/10.4324/9780203852279

Hughes, C. (1998). Executive function in preschoolers: Links with theory of mind and verbal ability. *British Journal of Developmental Psychology, 16*(2), 233-253. https://doi.org/10.1111/j.2044-835X.1998.tb00921.x

Hughes, C., Ensor, R., Wilson, A., & Graham, A. (2010). Tracking executive function across the transition to school: a latent variable approach. *Developmental Neuropsychology*, *35*(1), 20–36. https://doi.org/10.1080/87565640903325691

Kline, P. (1994). An easy guide to factor analysis. New York, NY: Routledge.

Lan, X., Legare, C. H., Ponitz, C. C., Li, S., & Morrison, F. J. (2011). Investigating the links between the subcomponents of executive function and academic achievement: A cross-cultural analysis of Chinese and American preschoolers. *Journal of Experimental Child*

Psychology, *108*(3), 677–692. https://doi.org/10.1016/j.jecp.2010.11.001

Lee, K., Bull, R., & Ho, R. M. H. (2013). Developmental changes in executive functioning. *Child Development*, *84*(6), 1933–1953. https://doi.org/10.1111/cdev.12096

Lehto, J. E., Juujärvi, P., Kooistra, L., & Pulkkinen, L. (2003). Dimensions of executive functioning: Evidence from children. British Journal of Developmental Psychology, 21(1), 59–80. https://doi.org/10.1348/026151003321164627

Lenes, R., Gonzales, C. R., Størksen, I., & McClelland, M. M. (2020). Children's Self-Regulation in Norway and the United States: The Role of Mother's Education and Child Gender Across Cultural Contexts. *Frontiers in Psychology*, *11* https://doi.org/10.3389/fpsyg.2020.566208

McClelland, M. M., Acock, A. C., & Morrison, F. J. (2006). The impact of kindergarten learning-related skills on academic trajectories at the end of elementary school. *Early Childhood Research Quarterly*, *21*(4), 471–490. https://doi.org/10.1016/j.ecresq.2006.09.003

McClelland, M. M., Acock, A. C., Piccinin, A., Rhea, S. A., & Stallings, M. C. (2013). Relations between Preschool Attention Span-Persistence and Age 25 Educational Outcomes. *Early Childhood Research Quarterly*, *28*(2), 314–324.https://doi.org/10.1016/j.ecresq.2012.07.008

McClelland, M. M., & Cameron, C. E. (2012). Self-Regulation Early Childhood: Improving Conceptual Clarity and Developing Ecologically Valid Measures. *Child Development Perspectives*, *6*(2), 136–142. https://doi.org/10.1111/j.1750-8606.2011.00191.x

McClelland, M. M., Cameron, C. E., Connor, C. M. D., Farris, C. L., Jewkes, A. M., & Morrison, F. J. (2007). Links Between Behavioral Regulation and Preschoolers' Literacy, Vocabulary, and Math Skills. *Developmental Psychology*, *43*(4), 947–959.

https://doi.org/10.1037/0012-1649.43.4.947

McClelland, M. M., Cameron, C. E., Duncan, R., Bowles, R. P., Acock, A. C., Miao, A., & Pratt, M. E. (2014). Predictors of early growth in academic achievement: The head-toes-knees-shoulders task. *Frontiers in Psychology*, *5*, 1–14. https://doi.org/10.3389/fpsyg.2014.00599

McClelland, M. M., Cameron Ponitz, C., Messersmith, E., & Tominey, S. (2010). Self-regulation: The integration of cognition and emotion. In W. F. Overton (Vol. Ed.) & R. Lerner (Eds.), *Handbook of life-span human development. Vol. 1: Cognition, biology and methods* (pp. 509-553). Hoboken, NJ: Wiley and Sons.

McClelland, M. M., Cameron, C. E., Wanless, S. B., & Murray, A. (2007). Executive function, behavioral self-regulation, and social-emotional competence: Links to school readiness. In O. N. Saracho & B. Spodek (Eds.), *Contemporary perspectives on social learning in early childhood education* (pp. 83-107). Charlotte, NC: Information Age.

McClelland, M. M., Geldhof, J., Cameron, C. E., & Wanless, S. B. (2015). Development and self-regulation. In W. F. Overton & P. C. M. Molenaar (Eds.), *Handbook of child psychology and developmental science* (7th ed., Vol. 1, pp. 523-565). Hoboken, NJ: Wiley.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27*(1), 82–99. https://doi.org/10.1111/j.2044-8317.1974.tb00530.x

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/10.1006/cogp.1999.0734

Miyake, A., & Friedman, N. P. (2012). The Nature and Organization of Individual Differences in

Executive Functions: Four General Conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14. https://doi.org/10.1177/0963721411429458

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., … Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(7), 2693–2698. https://doi.org/10.1073/pnas.1010076108

Morrison, F. J., Ponitz, C. C., & McClelland, M. M. (2010). Self-regulation and academic achievement in the transition to school. In S. D. Calkins & M. Bell (Eds.), *Child development at the intersection of emotion and cognition.* (pp. 203-224): Washington, DC, US: American Psychological Association.

Morrison, F. J., & Grammer, J. K. (2016). Conceptual clutter and measurement mayhem: Proposals for cross-disciplinary integration in conceptualizing and measuring executive function. In J. A. Griffin (Ed.), *Executive function in preschool-age children: Integrating measurement, neurodevelopment, and translational research , (pp* (Vol. 362, pp. 327–348). https://doi.org/10.1037/14797-015

Morrow, V., & Richards, M. (1996). The ethics of social research with children: An overview. *Children & Society, 10*(2), 90-105. https://doi.org/10.1111/j.1099-0860.1996.tb00461.x

Muñoz-Sandoval, A. F., Woodcock, R. W., McGrew, K. S., and Mather, N. (2005). *Batería III Woodcock- Muñoz: Pruebas de Aprovechamiento*. Itasca, IL: Riverside Publishing.

Muthén, L.K. and Muthén, B.O. (1998-2017). *Mplus User's Guide*. Eighth Edition. Los Angeles, CA: Muthén & Muthén

Muthén, B., du Toit, S.H.C. & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished technical report. Retrieved from https://www.statmodel.com/download/Article_075.pdf

Pears, K. C., Kim, H. K., Healey, C. V., Yoerger, K., & Fisher, P. A. (2015). Improving child self-regulation and parenting in families of pre-kindergarten children with developmental disabilities and behavioral difficulties. *Prevention Science: The Official Journal of the Society for Prevention Research*, *16*(2), 222–232. https://doi.org/10.1007/s11121-014-0482-2

Posner, M. I., & Rothbart, M. K. (1998). Attention, self-regulation and consciousness. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *353*(1377), 1915–1927. https://doi.org/10.1098/rstb.1998.0344

Purpura, D. J., Schmitt, S. A., & Ganley, C.M. (2017). Foundations of mathematics and literacy: The role of executive functioning components. *Journal of Experimental Child Psychology, 153*, 15-34. https://doi.org/10.1016/j.jecp.2016.08.010

Rainelli, S., Bulotsky-Shearer, R. J., Fernandez, V. A., Greenfield, D. B., & López, M. (2017). Validity of the first two subtests of the preschool language assessment scale as a language screener for Spanish-speaking preschool children. *Early Childhood Research Quarterly, 38*, 10-22. https://doi.org/10.1016/j.ecresq.2016.08.001

Rennie, D. A. C., Bull, R., & Diamond, A. (2004). Executive functioning in preschoolers: Reducing the inhibitory demands of the dimensional change card sort task. *Developmental Neuropsychology, 26*, 423-443. https://doi.org/10.1207/s15326942dn2601_4

Rhoades, B. L., Greenberg, M. T., & Domitrovich, C. E. (2009). The contribution of inhibitory control to preschoolers' social–emotional competence. *Journal of Applied Developmental Psychology*, *30*(3), 310–320. https://doi.org/10.1016/j.appdev.2008.12.012

Rueda, M. R., Posner, M. I., & Rothbart, M. K. (2005). The development of executive attention: contributions to the emergence of self-regulation. *Developmental Neuropsychology*, *28*(2), 573–594. https://doi.org/10.1207/s15326942dn2802_2

Schmitt, S. A., Geldhof, G. J., Purpura, D. J., Duncan, R., & McClelland, M. M. (2017). Examining the relations between executive function, math, and literacy during the transition to kindergarten: A multi-analytic approach. *Journal of Educational Psychology*, *109*(8), 1120–1140. https://doi.org/10.1037/edu0000193

Schmitt, S. A., Pratt, M. E., & McClelland, M. M. (2014). Examining the validity of behavioral self-regulation tools in predicting preschoolers' academic achievement. *Early Education and Development, 25*(5), 641-660. https://doi.org/10.1080/10409289.2014.850397

Schrank, F. A., McGrew, K. S., Ruef, M. L., Alvarado, C. G., Muñoz-Sandoval, A. F., & Woodcock, R. W. (2005). *Overview and technical supplement* (Bateria III Woodcock-Muñoz Assessment Service Bulletin No. 1). Itasca, IL: Riverside.

Senn, T. E., Espy, K. A., & Kaufmann, P. M. (2004). Using path analysis to understand executive function organization in preschool children. *Developmental Neuropsychology*, *26*(1), 445–464. https://doi.org/10.1207/s15326942dn2601_5

Soska, K. C., Adolph, K. E., & Johnson, S. P. (2010). Systems in development: Motor skill acquisition facilitates three-dimensional object completion. *Developmental Psychology, 46*(1), 129-138. https://doi.org/10.1037/a0014618

StataCorp. (2018). *Stata Statistical Software: Release 15.1*. College Station, TX: StataCorp LP.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Størksen, I., Ellingsen, I. T., Wanless, S. B., & McClelland, M. M. (2015). The Influence of

Parental Socioeconomic Background and Gender on Self-Regulation Among 5-Year-Old Children in Norway. *Early Education and Development*, *26*(5-6), 663–684. https://doi.org/10.1080/10409289.2014.932238

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(1), 104–132. https://doi.org/10.1037/0033-295X.114.1.104

von Suchodoletz, A., Gestsdottir, S., Wanless, S. B., McClelland, M. M., Birgisdottir, F., Gunzenhauser, C., & Ragnarsdottir, H. (2013). Behavioral self-regulation and relations to emergent academic skills among children in Germany and Iceland. *Early Childhood Research Quarterly*, *28*(1), 62–73. https://doi.org/10.1016/j.ecresq.2012.05.003

Wanless, S. B., McClelland, M. M., Acock, A. C., Ponitz, C. C., Son, S.-H., Lan, X., … Li, S. (2011). Measuring behavioral regulation in four societies. *Psychological Assessment*, *23*(2), 364–378. https://doi.org/10.1037/a0021768

Wanless, S. B., McClelland, M. M., Lan, X., Son, S.-H., Cameron, C. E., Morrison, F. J., … Sung, M. (2013). Gender differences in behavioral regulation in four societies: The United States, Taiwan, South Korea, and China. *Early Childhood Research Quarterly*, *28*(3), 621–633. https://doi.org/10.1016/j.ecresq.2013.04.002

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. Handbook of Structural Equation Modeling, 209–231.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods,12*(1), 58–70. https://doi.org/10.37/1082-989X.12.1.58

Wiebe, S. A., Sheffield, T., Nelson, J. M., Clark, C. A. C., Chevalier, N., & Espy, K. A. (2011).

The structure of executive function in 3-year-olds. *Journal of Experimental Child Psychology*, *108*(3), 436–452. https://doi.org/10.1016/j.jecp.2010.08.008

Willoughby, M. T., Blair, C. B., Wirth, R. J., & Greenberg, M. (2010). The measurement of executive function at age 3 years: Psychometric properties and criterion validity of a new battery of tasks. *Psychological Assessment, 22*(2), 306-317. https://doi.org/10.1037/a0018708

Willoughby, M. T., Blair, C. B., Wirth, R. J., & Greenberg, M. (2012). The measurement of executive function at age 5: Psychometric properties and relationship to academic achievement. *Psychological Assessment*, *24*(1): 226-239. https://doi.org/10.1037/a0025361

Willoughby, M. T., Wirth, R. J., & Blair, C. B. (2012). Executive function in early childhood: Longitudinal measurement invariance and developmental change. *Psychological Assessment, 24*(2), 418-431. https://doi.org/10.1037/a0025779

Woodcock, R. W., & Mather, N. (2000). *Woodcock Johnson Psycho-Educational Battery-III.* Itasca, IL: Riverside.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Examiner's Manual. Woodcock-Johnson III Tests of Achievement.* Itasca, IL: Riverside Publishing.

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature protocols*, *1*(1), 297-301 https://doi.org/10.1038/nprot.2006.46

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

Zelazo, P. D., & Cunningham, W. A. (2007). Executive Function: Mechanisms Underlying Emotion Regulation. In J. J. Gross (Ed.), *Handbook of emotion regulation , (pp* (Vol. 654, pp. 135–158). Retrieved from https://psycnet.apa.org/fulltext/2007-01392-007.pdf

Table 1

Sample demographics and descriptive statistics of self-regulation measures

	Fall preschool		Spring Preschool		Fall Kindergarten		Spring Kindergarten	
	N	M (SD)	N	M (SD)	N	M (SD)	N	M (SD)
Age	172	56.10 (3.81)	151	61.50 (3.77)	133	67.54 (3.86)	126	73.56 (3.73)
Percent Female	172	54%	151	56%	133	54%	126	54%
Percent ELL	172	16%	151	8%	133	6%	126	6%
HTKS-R	169	36.81 (26.89)	137	53.51 (30.23)	133	75.35 (30.45)	126	88.59 (25.05)
Day-Night	167	20.39 (9.44)	139	23.07 (8.81)	133	26.28 (6.68)	126	28.7 (4.88)
DCCS	163	10.87 (5.67)	131	13.55 (5.60)	132	16.00 (4.90)	126	17.53 (4.02)
Working Memory	160	445.28 (10.57)	128	450.23 (16.25)	133	449.93 (17.79)	126	463.23 (19.57)

Table 2

Distributional properties of scores on the Head-Toes-Knees-Shoulders-Revised Task (Time 1)

Age Group	N	Mean	Skewness	Kurtosis	HTKS-R % Floor	HTKS % Floor
48 Months	32	27.74 (20.41)	1.61	5.60	3.23%	22.58%
52 Months	51	34.63 (25.61)	1.01	3.23	1.96%	17.65%
56 Months	55	43.98 (29.24)	0.57	2.16	3.64%	12.73%
60 Months	33	36.75 (28.04)	1.09	3.05	3.12%	15.62%
Overall	169	36.81 (26.89)	0.97	2.95	2.96%	16.57%

Table 3

Correlations, means, and standard deviations of each subpart of the HTKS-R across the four time points.

	Fall Preschool				Spring Preschool				Fall Kindergarten				Spring Kindergarten			
	P0	P1	P2	P3	P0	P1	P2	P3	P0	P1	P2	P3	P0	P1	P2	P3
Part 0																
Part 1	.48***				.38***				.47***				.52***			
Part 2	.34***	.80***			.28***	.82***			.39***	.81***			.52***	.78***		
Part 3	.27***	.62***	.83***		.16	.56***	.81***		.26**	.57***	.73***		.29***	.45***	.66***	
Mean	17.54	11.10	5.38	2.70	19.61	17.52	10.27	6.10	21.11	25.37	17.41	11.63	21.17	29.29	23.01	15.11
SD	6.24	10.61	8.74	6.59	4.77	11.64	10.53	9.19	3.05	10.50	11.05	11.05	2.71	7.19	8.94	11.00

** $p < .01$ *** $p < .001$
Note: Maximum possible score is 22 points in Part 0, 34 points in Part 1, 30 points in Part 2, and 32 points in Part 3.

Table 4

Results of Item Factor Analyses (IFA) of HTKS-R across all observations

Mean (and Range) of Rotated Factor Loadings of each Subsection of the HTSK-R by each Result.						
	1 Factor Result	2 Factor Result		3 Factor Result		
	Factor 1	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
Part 0	.78 (.63 * : .92 *)	.91 (.75 * : 1.05 *)	-.03 (-.18 : .07)	.88 (.71* : 1.05*)	.07 (-.10 : .20*)	-.04 (-.23 : .16)
Part 1	.96 (.93 *: .99 *)	.52 (.43 * : .63 *)	.60 (.49 * : .67 *)	.00 (-.08 : .05)	.83 (.76* : .88*)	.38 (.29* : .50*)
Part 2	.94 (.88 * : .97 *)	.30 (.21 * : .35 *)	.76 (.66 * : .85 *)	.24 (.14* : .24*)	.78 (.70* : .85*)	.06 (-.05 : .14)
Part 3	.90 (.88 * : .96 *)	-.05 (-.17 * : .10)	.96 (.87 * : 1.06 *)	-.12 (-.40* : .40*)	1.01 (.94* : 1.18*)	-.13 (-.28* : .00)

Model Fit Information for each Result			
	1 Factor Result (df = 629)	2 Factor Result (df = 593)	3 Factor Result (df = 558)
χ^2	1789.56	1095.08	781.11
RMSEA	.057	.039	.027
SRMR	.113	.083	.059
CFI	.991	.996	.998
TLI	.990	.996	.998

Table 5

Pairwise correlations between demographic variables and measures of self-regulation at each timepoint

Variable	1	2	3	4	5	6	7	8	9	10	11
1. Age	-	-.03	.08	.26 **	.20 *	.11	-.01	.12	.15	.10	.16 †
2. Gender (male = 1)	-.03	-	.21 **	.01	.00	.04	-.08	-.03	.10	-.12	-.22 *
3. ELL Status	.08	.21 **	-	-.17 †	.01	-.41 ***	-.17 †	-.12	.02	-.30 ***	-.14
4. Fall HTKS-R	.17 *	-.10	-.10	-	.44 ***	.34 ***	.23 **	.60 ***	.33 ***	.36 ***	.41 ***
5. Fall Day Night	.05	.06	.18	.26 ***	-	.25 **	.13	.32 ***	.51 ***	.21 *	.20 *
6. Fall DCCS	.14 †	-.06	-.03	.43 ***	.10	-	.15 †	.34 ***	.28 **	.41 ***	.26 **
7. Fall Working Memory	.10	-.05	-.29	.19 *	.15 †	.17 *	-	.23 *	.11	.08	.44 ***
8. Spring HTKS-R	.27 **	-.09	-.09	.69 ***	.31 ***	.38 ***	.26 **	-	.31 ***	.31 ***	.36 ***
9. Spring Day Night	.15	-.10	.12	.24 **	.35 ***	.25 **	.18 *	.27 **	-	.16 †	.17 †
10. Spring DCCS	.25 **	-.18	-.17 †	.44 ***	.04	.32 ***	.20 *	.43 ***	.27 **	-	.31 ***
11. Spring Working Memory	.00	-.06	-.18 *	.31 ***	.21 *	.22 *	.22 *	.39 ***	.19 *	.33 ***	-

† $p < .10$ * $p < .05$ ** $p < .01$ *** $p < .001$

Note: the bottom diagonal contains data from Times 1 and 2 in the preschool year, and the Top diagonal contains data from Times 3 and 4 in the kindergarten year.

Table 6

Pairwise correlations between HTKS and HTKS-R with other measures of self-regulation at each timepoint.

	Fall Preschool				Spring Preschool			
	HTKS	HTKS-R	z	*p*	HTKS	HTKS-R	z	*p*
HTKS-R	.96***	-	-	-	.97***	-	-	-
Day- Night	.22**	.27***	1.833	.033	.22*	.28**	2.050	.020
DCCS	.41***	.43***	0.699	.242	.40***	.43***	0.908	.182
Working Mem.	.18*	.19*	0.513	.304	.39***	.39***	0.047	.481
	Fall Kindergarten				**Spring Kindergarten**			
	HTKS	HTKS-R	z	*p*	HTKS	HTKS-R	z	*p*
HTKS-R	.98***	-	-	-	.98***	-	-	-
Day- Night	.41***	.44***	1.974	.024	.24**	.31***	4.424	< .001
DCCS	.29***	.34***	3.07	.001	.29**	.32***	1.65	.049
Working Mem.	.23**	.23**	0.094	.463	.34***	.36***	1.44	.075
	Fall Preschool to Fall Kindergarten				**Spring Preschool to Spring Kindergarten**			
	HTKS	HTKS-R	z	*p*	HTKS	HTKS-R	z	*p*
Day- Night	.25**	.34***	2.939	.002	.31***	.38***	2.605	.005
DCCS	.23*	.32**	3.107	.001	.22**	.27**	1.577	.057
Working Mem.	.45***	.46***	0.159	.437	.59***	.61***	0.811	.209

Note: Z tests and corresponding p-values test for significant differences between the HTKS and HTKS-R for each of the EF measures.
* p < .05 ** p < .01 *** p < .001

Supplemental Materials

Below, we provide results of the underpowered item factor analyses which evaluates the communality of individual items and underlying factor structure with data from each of the individual time points as described in the main body of the study. For each time point, we evaluated a one-factor, two-factor, and three-factor solution, and provide estimates of absolute and relative model fit for each model and provide model-fit comparison tests to evaluate whether there were significant increases in absolute model fit between each of the solutions within each timepoint. Given the underpowered nature of these exploratory analyses, results should be interpreted with caution.

Time 1 (Fall of Preschool, N = 169). Results for the IFA for the one-, two- and three-factor solutions for Time 1 are presented in Supplemental Table 1. As shown in the top of Supplemental Table 1, the pattern of standardized factor loadings across the three solutions followed the same general pattern as the in the overall model from the main body of the study. As shown in the bottom of Supplemental Table 1, all three solutions demonstrated acceptable to excellent levels of relative model fit with only small changes in relative model fit between all three solutions.

Factors within both the two- and three-factor solution were all significantly correlated with one another, and there were significant increases in absolute model fit between each of the solutions as well. In the two-factor solution, the first factor was significantly correlated to the second factor, $r = .411$, $p < .05$, and there was a significant increase in absolute model fit compared to the one-factor solution, χ^2 (36) = 105.26, $p < .001$. For the three-factor solution, the first factor was correlated with the second factor, $r = .403$, $p < .05$, and the second factor was also correlated with the third factor, $r = .331$, $p < .05$. There was also a significant increase in absolute model fit compared to the two-factor solution, χ^2 (35) = 55.99, $p = .014$.

Time 2 (Spring of Preschool, N = 137). Results for the IFA for the one-, two-, and three-factor solutions for Time 2 are presented in Supplemental Table 2. Like the previous time-specific model, the pattern of standardized factor loadings across the three solutions followed the same general pattern as the in the overall model, and all three solutions also demonstrated acceptable to excellent levels of relative model fit.

Patterns of factor correlations differed from models in the previous time-point, but there continued to be significant improvements in absolute model fit between each of the solutions. Unlike the models from the previous timepoint, the correlation between the first and second factor in the two factor solution was not significant, and in the three factor solution, only the correlation between the second and third factor was significant, $r = .687$, $p < .001$. Like in the previous timepoint, there was a significant increase in absolute model fit in the two-factor solution compared to the one-factor solution, χ^2 (36) = 128.52, $p < .001$. For the three-factor solution, there was also a significant increase in absolute model fit compared to the two-factor solution, χ^2 (35) = 78.05, $p < .001$.

Time 3 (Fall of Kindergarten, N = 133). Results for the IFA for the one-, two-, and three-factor solutions for Time 3 are presented in Supplemental Table 3. Like the two-previous time specific models, the pattern of standardized factor loadings across all three solutions continued to show the same general pattern of standardized factor loadings. In terms of relative model-fit, all three solutions continued to show acceptable to excellent levels of relative model fit with only small changes between each of solutions.

Patterns of factor correlations within each solution in this time point were more similar to the first time point, but there continued to be significant improvements of absolute model fit between the three solutions. In the two-factor solution, the first factor was significantly correlated to the second factor, $r = .389$, $p < .05$, and there was a significant increase in absolute model fit compared to the one-factor solution, χ^2 (36) = 128.96, $p < .001$. For the three-factor solution, the first factor was correlated with the second factor, $r = .695$, $p < .05$, and the third factor, $r = .450$, $p < .05$. The second factor was also correlated with the third factor, $r = .633$, $p < .05$. There was also a significant increase in absolute model fit compared to the two-factor solution, χ^2 (35) = 74.48, $p < .001$.

Time 4 (Spring of Kindergarten, N = 127). Results for the IFA for the one-, two-, and three-factor solutions for Time 4 are presented in Supplemental Table 4. Patterns of standardized factor loadings across the three-solutions within this wave continued to follow the same general trend as previous models, and all three models showed acceptable to excellent levels of relative model fit. Although still small and within the range of excellent model fit, changes in CFI between the one and two-factor solution were the largest out of any of the previous solutions within the time-specific models.

Similar to the first two time-specific models, in the two factor solution, the first and second factor were significantly correlated with one another, $r = .422$, $p < .05$ and there was a significant improvement in absolute model fit compared to the one-factor solution, χ^2 (35) = 115.60, $p < .001$. In the three factor solution, only the first and third factor were significantly correlated with one another, $r = .586$, $p < .05$, but there was another significant improvement in terms of absolute model fit compared to the two-factor solution, , χ^2 (35) = 73.94, $p < .001$.

Supplemental Table 1

IFA Analyses of HTKS-R at Time 1 (Fall of Preschool, N = 169)

Mean (and Range) of Rotated Factor Loadings of each Subsection of the HTSK-R by each Result.

	1 Factor Result	2 Factor Result		3 Factor Result		
	Factor 1	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
Part 0	.78 (.59* : .92*)	.86 (.63* : .97*)	.03 (-.13 : .09)	.77 (.57* : .85*)	.05 (-.06 : .19)	.36 (-.01 : .63*)
Part 1	.93 (.84* : .99*)	.32 (.14 : .51*)	.76 (.62* : .87*)	.00 (-.17 : .11)	.77 (.66* : .89*)	.35 (-.12 : .58*)
Part 2	.94 (.88* : .98*)	.07 (.00 : .19*)	.92 (.80* : .98*)	-.16 (-.29* : .09)	.99 (.81* : 1.04*)	.03 (-.01 : .13)
Part 3	.93 (.88* : .98*)	-.22 (-.37* : -.07)	1.02 (.96 * : 1.07 *)	.06 (-.20 : .36*)	.97 (.77* : 1.09*)	-.22 (-.43* : .02*)

Model Fit Information for each Result

	1 Factor Result (df = 629)	2 Factor Result (df = 593)	3 Factor Result (df = 558)
χ^2	752.414	611.52	547.91
RMSEA	.034	.014	< .001
SRMR	.193	.103	.075
CFI	.991	.999	.999
TLI	.991	.999	.999

Supplemental Table 2

IFA Analyses of HTKS-R at Time 2 (Spring of Preschool, N = 137)

Mean (and Range) of Rotated Factor Loadings of each Subsection of the HTSK-R by each Result.

	1 Factor Result	2 Factor Result		3 Factor Result		
	Factor 1	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
Part 0	.70 (.44* : 1.04 *)	.91 (.82* : 1.01*)	-.06 (-.28 : .11)	.85 (.72* : .88*)	.13 (-.06 : .42*)	.13 (-.02 : .46)
Part 1	.95 (.89 *:.99 *)	.52 (.43* : .61*)	.68 (.57 * : .75*)	.10 (-.07 : .20)	.95 (.78* : 1.07*)	-.01 (-.14 : .20)
Part 2	.91 (.82* : .96*)	.07 (-.05 : .21*)	.90 (.76 * : .97*)	-.04 (-.27* : .06)	.43 (.33* : .57*)	.58 (.39* : .69*)
Part 3	.89 (.83 * : .94 *)	-.12 (-.28* : .11)	.94 (.86 * : .99*)	-.07 (-.46* : .29*)	.18 (-.15 : .63*)	.78 (.38* : 1.05*)

Model Fit Information for each Result

	1 Factor Result (df = 629)	2 Factor Result (df = 593)	3 Factor Result (df = 558)
χ^2	899.82	667.34	583.92
RMSEA	.056	.030	.018
SRMR	.235	.111	.081
CFI	.983	.995	.998
TLI	.982	.995	.998

Supplemental Table 3

IFA Analyses of HTKS-R at Time 3 (Fall of Kindergarten, N = 133)

Mean (and Range) of Rotated Factor Loadings of each Subsection of the HTSK-R by each Result.

	1 Factor Result	2 Factor Result		3 Factor Result		
	Factor 1	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
Part 0	.75 (.61* : .90*)	.93 (.74* : 1.00*)	-.03 (-.09 : .01)	1.07 (.99* : 1.10*)	-.07 (-.34* : .11)	-.27 (-.37* : -.19)
Part 1	.97 (.95 * : 1.00*)	.58 (.54* : .66*)	.59 (.52* : .64*)	.50 (.40* : .58*)	.52 (.38* : .71*)	.06 (-.13 : .27)
Part 2	.89 (.82* : .99*)	.41 (.28 : .57*)	.66 (.57* : .85*)	.03 (-.10 : .15)	.92 (.72* : 1.16*)	-.04 (.31* : .13*)
Part 3	.87 (.77 * : .99*)	-.05 (-.19* : .13)	.93 (.79* : 1.06*)	.11 (-.02 : .27*)	.25 (-.01 : .38*)	.68 (.50* : 88*)

Model Fit Information for each Result

	1 Factor Result (df = 629)	2 Factor Result (df = 593)	3 Factor Result (df = 558)
χ^2	881.38	656.91	567.24
RMSEA	.055	.028	.011
SRMR	.166	.081	.058
CFI	.983	.996	.999
TLI	.982	.995	.999

Supplemental Table 4

IFA Analyses of HTKS-R at Time 4 (Spring of Kindergarten, N = 127)

Mean (and Range) of Rotated Factor Loadings of each Subsection of the HTSK-R by each Result.

	1 Factor Result	2 Factor Result		3 Factor Result		
	Factor 1	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
Part 0	.70 (.14* : .95*)	.92 (.47* : 1.07*)	-.29 (-.45 : -.17)	.30 (-.05 : .74*)	.73 (.57* : .93*)	.14 (-.03 : .29)
Part 1	.92 (.83* : .99*)	.83 (.67* : .99*)	.20 (.00 : .39*)	.94 (.78* : 1.12*)	-.05 (-.17 : .06)	.01 (-.23 : .26)
Part 2	.91 (.75* : .99*)	.59 (.47* : .77*)	.51 (.37* : .66*)	.58 (.47* : .73*)	.10 (-.03 : .27*)	.44 (.24 : .61*)
Part 3	.88 (.79* : .97*)	.06 (-.10 : .22)	.89 (.76* : 1.00*)	.02 (-.15 : .14)	-.19 (-.48* : .08)	.92 (.78* : 1.06*)

Model Fit Information for each Result

	1 Factor Result (df = 629)	2 Factor Result (df = 593)	3 Factor Result (df = 558)
χ^2	844.87	666.14	566.63
RMSEA	.052	.031	.011
SRMR	.220	.143	.080
CFI	.978	.992	.999
TLI	.976	.991	.999