Received: 29 June 2021 | Revised: 5 April 2022 | Accepted: 22 April 2022

DOI: 10.1111/desc.13280

Developmental Science WILEY

RESEARCH ARTICLE

Studying children's growth in self-regulation using changing measures to account for heterotypic continuity: A Bayesian approach to developmental scaling

Alexis Hosch[1] | Jacob J. Oleson[2] | Jordan L. Harris[1] | Mary Taylor Goeltz[3] | Tabea Neumann[4] | Brandon LeBeau[5] | Eliot Hazeltine[1] | Isaac T. Petersen[1]

[1]Department of Psychological and Brain Sciences, University of Iowa, Iowa City, Iowa, USA

[2]Department of Biostatistics, University of Iowa, Iowa City, Iowa, USA

[3]Department of Psychology, University of Colorado Denver, Denver, Colorado, USA

[4]School of Medicine, University of Dundee, Dundee, Scotland, UK

[5]Department of Psychological and Quantitative Foundations, University of Iowa, Iowa City, Iowa, USA

Correspondence

Alexis Hosch, Department of Psychological and Brain Sciences, University of Iowa, G60 Psychological and Brain Sciences Building, Iowa City, IA 52242, USA.
Email: alexis-hosch@uiowa.edu

Funding information

Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), Grant/Award Number: HD098235; National Institute of General Medical Sciences (NIGMS), Grant/Award Number: T32GM108540; National Center for Advancing Translational Sciences (NCATS), Grant/Award Number: UL1TR002537

Abstract

Self-regulation is thought to show heterotypic continuity–its individual differences endure but its behavioral manifestations change across development. Thus, different measures across time may be necessary to account for heterotypic continuity of self-regulation. This longitudinal study examined children's (N = 108) self-regulation development using 17 measures, including 15 performance-based measures, two questionnaires, and three raters across seven time points. It is the first to use different measures of self-regulation over time to account for heterotypic continuity while using developmental scaling to link the measures onto the same scale for more accurate growth estimates. Assessed facets included inhibitory control, delayed gratification, sustained attention, and executive functions. Some measures differed across ages to retain construct validity and account for heterotypic continuity. A Bayesian longitudinal mixed model for developmental scaling was developed to link the differing measures onto the same scale. This allowed charting children's self-regulation growth across ages 3–7 years and relating it to both predictors and outcomes. Rapid growth occurred from ages 3–6. As a validation of the developmental scaling approach, greater self-regulation was associated with better school readiness (math and reading skills) and fewer externalizing problems. Our multi-wave, multi-facet, multi-method, multi-measure, multi-rater, developmental scaling approach is the most comprehensive to date for assessing the development of self-regulation. This approach demonstrates that developmental scaling may enable studying development of self-regulation across the lifespan.

KEYWORDS

changing measures, construct validity invariance, developmental scaling, heterotypic continuity, longitudinal, self-regulation

The present study is part of a larger study, the School Readiness Study. Measures and hypotheses for the School Readiness Study were pre-registered: https://osf.io/jzxb8.

1 | INTRODUCTION

Considerable research has demonstrated that children's ability to willfully regulate their thoughts, emotions, and behaviors holds important implications for their long-term outcomes. This ability, commonly referred to as self-regulation, involves the flexible control of attentional, cognitive, emotional, and behavioral processes in pursuit of a goal (Berger, 2011; Calkins & Fox, 2002). Self-regulation has shown concurrent and predictive associations with myriad outcomes in childhood, including internalizing and externalizing problems (Eisenberg et al., 2009; Espy et al., 2011; Martel & Nigg, 2006; Olson et al., 2005; Petitclerc et al., 2015; Rothbart & Bates, 2006), social and intellectual functioning (Blair & Razza, 2007; Kochanska, 1997; Padilla-Walker & Christensen, 2011; Spinrad et al., 2006), and school readiness (Blair & Diamond, 2008; Liew, 2012; Liew et al., 2018). Furthermore, longitudinal research has shown that children with poorer self-regulation tend to have worse health, less wealth, and more criminal involvement as adults (Moffitt et al., 2011). Thus, it is crucial to investigate how self-regulation develops.

Self-regulation has been conceptualized in many ways, due partly to a lack of consensus regarding accepted terminology and to differing emphases of various research traditions. For example, neuropsychologists have examined self-regulatory processes called executive functions—higher-order ("top-down") processes that exert control over attentional, cognitive, and behavioral tendencies in pursuit of a goal (Zhou et al., 2012). Temperament researchers have proposed that self-regulation is the result of an executive attentional control system (Shallice, 1988), in which effortful control reflects the efficiency of the executive attentional control system's ability to inhibit prepotent responses, plan behavior, and detect errors (Posner & Rothbart, 2000; Rothbart & Bates, 2006; Tiego et al., 2020). Other developmental researchers have argued that there should be greater consideration of emotional processes within a framework of self-regulation (Cole et al., 2004; Eisenberg et al., 2001; Lewis & Stieben, 2004; Mischel & Ayduk, 2004). Consequently, considerable work has examined self-regulation as a regulatory system that involves distinct "hot" (i.e., motivationally or affectively mediated) and "cool" (i.e., cognitively mediated) processes (Backer-Grøndahl et al., 2019; Bechara et al., 1994; Cameron Ponitz et al., 2008; Denham et al., 2012; Metcalfe & Mischel, 1999; Simpson & Carroll, 2019; Willoughby et al., 2011). For instance, some factor analysis research has demonstrated that tasks designed to assess the inhibitory control aspect of self-regulation—the ability to inhibit responses to irrelevant stimuli in pursuit of a cognitively represented goal—load onto separate latent "hot" and "cool" factors (Bridgett et al., 2015; Carlson & Moses, 2001; Murray & Kochanska, 2002; Simpson & Carroll, 2019).

The "hot" and "cool" factor conceptualization of self-regulation is not universally supported, however. Some have argued that an integrated, single factor model of regulation more accurately represents the construct, particularly in early childhood (Allan & Lonigan, 2011; Cole et al., 2019; Lin et al., 2019; Sulik et al., 2010; Wiebe et al., 2008). Thus, researchers have called for an integrated model of self-regulation, highlighting meaningful conceptual and measurement overlap between regulation-related constructs, including effortful control and executive function (Bridgett et al., 2013; Zhou et al., 2012), metacognition and executive function (Roebers, 2017), executive function and self-regulation (Best et al., 2009; Hofmann et al., 2012; McCoy, 2019; Roebers, 2017), executive function and emotion regulation (Zelazo & Cunningham, 2007), and cognitive control and self-regulation (Mischel et al., 2011).

In response to calls for an integrated model of self-regulation, Nigg (2017) proposed a domain-general conceptualization of self-regulation that integrates diverse constructs, across emotion, action, and cognition, into a unified framework to provide consistency and prevent confusion within the field. This domain-general framework may also aid the development and improvement of measurement techniques and interpretation of results. We draw upon Nigg's (2017) framework in our conceptualization of self-regulation and apply a domain-general model that encompasses various regulatory processes. We acknowledge that this is one of several empirically supported conceptualizations of self-regulation. Other approaches, such as a formative model of self-regulation, in which the construct is derived from the summation of a set of processes, may also reasonably operationalize the construct (Camerota et al., 2020; Willoughby et al., 2017). Alternatively, it is possible that what researchers have called "self-regulation" is merely a heuristic that describes a set of separate yet correlated abilities that do not reflect a common construct (Eisenberg et al., 2018, 2019). In sum, the structure of self-regulation is highly debated and remains an important empirical question. Research has supported several conceptualizations of self-regulation, including reflective, formative, or heuristic models. More research is needed to delineate how self-regulatory processes (e.g., inhibitory control, executive functions, etc.) are related within a broader developmental framework.

However, prior research has generally indicated that regulatory processes are inter-correlated and that there is considerable conceptual and measurement overlap between several components of self-regulation, including effortful control, executive functions, inhibitory

RESEARCH HIGHLIGHTS

- Self-regulation shows heterotypic continuity, but studies have not accounted for it when examining individuals' growth.
- This study used different measures across ages and developmental scaling to account for heterotypic continuity and chart children's self-regulation growth across ages 3–7 years.
- The developmentally scaled model of self-regulation placed age-differing measures onto the same scale and showed criterion validity in relation to externalizing problems and school readiness.
- Developmental scaling may promote studying individuals' self-regulation development across the lifespan.

control, and others (Berger, 2011; Bridgett et al., 2013; Carlson & Wang, 2007; Lin et al., 2019; Nigg, 2017; Reed et al., 2020; Zelazo & Cunningham, 2007; Zhou et al., 2012). Given the overlap between concepts, evidence suggests that there may be a general, over-arching factor (Allan & Lonigan, 2011; Espy et al., 2011; Lin et al., 2019; Sulik et al., 2010; Wiebe et al., 2008, 2011). Thus to prevent confusion across constructs and to facilitate more efficient communication across research groups (Cole et al., 2019; McClelland et al., 2010; Nigg, 2017; Zhou et al., 2012), we conceptualized self-regulation as a higher-order construct, reflecting cognitively and affectively mediated regulatory abilities, consistent with prior studies (Allan & Lonigan, 2011; Espy et al., 2011; Sulik et al., 2010; Wiebe et al., 2011). Cognitive regulatory processes include, for instance, sustained attention (i.e., the ability to maintain focus on a given task over prolonged periods), inhibitory control, and higher-order executive functions. Affective regulatory processes include, for instance, the ability to delay gratification (i.e., the ability to resist temptation in favor of long-term goals) and regulate emotions (Bridgett et al., 2015; Gagne et al., 2021; Metcalfe & Mischel, 1999).

Previous research has also not precisely delineated how regulatory abilities develop across the lifespan. Prior literature generally supports a developmental model in which lower-level processes develop in early childhood and are followed by higher-level processes in later childhood. Montroy et al. (2016) described a hierarchical differentiation framework, in which children develop separate skills that enable self-regulation in infancy and later integrate these processes into a hierarchically organized regulatory system. Indeed, research has shown that there is a qualitative shift in regulatory skills beginning at age three, in which rapid growth occurs and then shows marked deceleration around age seven (Cameron Ponitz et al., 2008; Diamond, 2002; Montroy et al., 2016; Wiebe et al., 2011). Similarly, inhibitory control and working memory processes manifest in the first years of life and increase in capacity across the preschool years (Geeraerts et al., 2021; Greene, 2017; Kopp, 1982). As children enter formal schooling around age five, their executive function capacity increases, which supports early manifestation of higher-order processes such as cognitive flexibility and active self-regulation of cognition, emotion, and behavior (Anderson, 2002; Berger, 2011; Greene, 2017).

Individual trajectories may differ from this prototypical developmental timeline. For example, Montroy et al. (2016) examined 1386 children aged 3–7 using an inhibitory control task and found that, while most children demonstrated a pattern of rapid development of self-regulation followed by a deceleration period, child-specific factors (e.g., gender and language ability) predicted when and how quickly this growth occurred. Thus, behavioral manifestations of self-regulation may change across development due to non-linear development of self-regulation, as well as child-specific individual differences.

Persistence of a construct, such as self-regulation, with behavioral manifestations that change across development is called heterotypic continuity (Cicchetti & Rogosch, 2002). Self-regulation is thought to manifest differently in preschool-aged children compared to those in later childhood (Greene, 2017; Kopp, 1982; Petersen et al., 2016). In early childhood, self-regulation is thought to reflect a gradual transition from external sources of control to internal self-control (Berger, 2011; Kopp, 1982). Infants are reliant on caregivers to provide regulation, such as soothing through feeding, diaper changing, or holding (Kopp, 1982). Infants are also able to reduce excessive arousal or stimulation by turning away or self-soothing (Kopp, 1982). Between ages 2 and 3 years, children begin to develop more sophisticated forms of cognition, such as language and representational thinking, which allow them to act intentionally and comply with external commands (Berger, 2011; Kochanska, 2002; Kochanska et al., 2001; Kopp, 1982). However, 2- and 3-year-old children are still largely reliant on caregivers and more likely to react with physical aggression and have emotional outbursts during this time (Kopp, 1982). Kopp refers to this phase as "self-control", a more limited form of self-regulation, characterized by the development of autonomy and self-awareness. "Real" forms of self-regulation begin to emerge between 3 and 4 years of age, in which children become increasingly able to use rules, strategies, and plans to guide behavior (Berger, 2011; Kopp, 1982). During this time, children may use private (self-directed) speech to guide thoughts and actions during challenging tasks (Berger, 2011; Berk, 1999; Bivens & Berk, 1990). Initially, private speech functions as a planning instrument, occurring before action, in which children regulate their actions verbally. Eventually, private speech is thought to become internalized between ages 6 and 8 years, and it serves as an internal regulatory mechanism (Berger, 2011; Berk, 1999; Bivens & Berk, 1990). Internalized private speech is considered critical for self-regulation (Berger, 2011). Language achievements and concomitant growth in self-control (Whedon et al., 2021) are paralleled by development in the prefrontal cortex (e.g., anterior cingulate cortex and dorsolateral prefrontal cortex) and executive functions. This growth, which typically occurs between ages 3 and 7 years, supports more sophisticated forms of self-regulation as children get older (Berger, 2011; Diamond, 2002; McClelland et al., 2010).

With age, an increase in developmental capacity, paired with environmental changes (e.g., school entry), leads to heterogeneous manifestations of self-regulation. For instance, verbally requesting a toy rather than employing an automatic response, such as physical aggression, may indicate overt self-regulation in younger children, whereas similar behavior in older childhood may not reflect the same degree of inhibition. Among older children, self-regulation may instead appear as inhibition of a prepotent behavioral response despite a concrete command (e.g., "Simon Says") or social pressure (e.g., invitation by a peer to participate in a rule-breaking action), or as completion of a homework assignment that requires integration of planning, working memory, and control. In general, elementary school children tend to be more responsible and conscious of their behavior compared to preschool children (Berger, 2011).

Empirical work supports the notion that self-regulation shows heterotypic continuity. Studies have examined the heterotypic continuity of specific components of self-regulation, including inhibitory control (Geeraerts et al., 2021; Petersen et al., 2016; Petersen, Bates, et al., 2021), effortful control (Putnam et al., 2008), and emotional/behavioral control (Chang et al., 2015; Zimmermann & Iwanski, 2014). However, no studies have examined the heterotypic continuity of the higher-order self-regulation construct. A meta-analysis found that the behavioral manifestations of inhibitory control changed across time in children between 1 and 8 years of age (Petersen et al., 2016). More specifically, findings suggested that perceptual inhibition may develop earlier than other forms of inhibition, such as performance and association inhibition, which in turn may develop earlier than motivational inhibition. Similar patterns have also been found for other components of self-regulation. For example, Chang et al. (2015) found that children displayed different forms of emotional and behavioral control in early childhood, in which the inability to master early regulatory skills hindered the development of more advanced regulatory skills. In summary, theoretical and empirical evidence suggests that self-regulation exhibits heterotypic continuity. That is, behavioral manifestations of self-regulation change across development despite the persistence of the construct. However, whether self-regulation demonstrates heterotypic continuity is ultimately an empirical question, and limited empirical work has tested this possibility. Previous empirical work examining this question has been limited to specific components of self-regulation. It is thus important for empirical studies to investigate whether the broader self-regulation construct demonstrates heterotypic continuity.

If self-regulation shows heterotypic continuity, there are important measurement implications. Using the same measure across development may not reflect the same construct at different ages (Widaman et al., 2010). That is, a given measure may not be developmentally appropriate or construct-valid at all ages, such that scores on the same measure across time may reflect differences in the measure's meaning, rather than real change in an individual's level of self-regulation (Petersen et al., 2016). Consequently, accounting for heterotypic continuity of self-regulation may require using different measures across time (Widaman et al., 2010), because children are expected to display different behaviors at different ages for the same underlying construct of self-regulation (Bates & Novosad, 2005). Studies examining children's self-regulation development should account for these changes by using different measures across ages (Petersen et al., 2016, 2020). Using different, age-relevant measures over time provides more accurate growth estimates, at the group- and person-level than approaches that ignore heterotypic continuity (Chen & Jaffee, 2015; Petersen et al., 2018; Petersen, LeBeau, et al., 2021). Although considerable research has examined different measures of self-regulation at different ages in recognition of its heterotypic continuity (e.g., Chang et al., 2015), no prior work has examined individuals' self-regulation *growth* using different measures across development. Thus, we use developmental scaling to estimate children's growth to better understand self-regulation development.

1.1 | The present study

In the present study, we apply a domain-general hierarchical model of self-regulation that includes multiple related regulatory processes: inhibitory control, delayed gratification, sustained attention, and executive functions, consistent with Berger (2011), McClelland et al. (2010), Nigg (2017), and others. Given task impurity (McCoy, 2019), we use multiple measures and assessment methods (i.e., performance-based measures and questionnaires) for more robust estimates of children's self-regulation (Gagne et al., 2021). Measures were chosen because they reflect a broad range of regulatory skills, including both lower-level and higher-level processes. Moreover, these measures are commonly used in studies examining development of self-regulation and have shown reliability and validity within the age range of the present study (Carlson, 2005; Petersen et al., 2016). To account for heterotypic continuity, we use some common measures across adjacent ages to capture the core self-regulation phenotype on the same scale, and some different measures across ages to capture the changing manifestation.

To date, no studies have developed a model to account for heterotypic continuity of self-regulation. In the present study, we use a Bayesian longitudinal mixed model for developmental scaling to link differing measures of self-regulation across ages onto the same scale. This model allows charting children's growth across ages 3–7 years. Bayesian longitudinal mixed modeling is ideal for this developmental scaling scenario. Bayesian implementation relies upon conditional logic for the model structure which simplifies how the model is framed and allows information to be borrowed across multiple measurements of people's standing on the latent self-regulation construct. The Bayesian model also leverages all available data to estimate latent growth curves. Furthermore, we have prior scientific knowledge of developmental scaling, and we apply this knowledge in the structure of the Bayesian model. By borrowing strength across the abundant data available per participant and by utilizing the informative prior structure, we can obtain reliable and interpretable estimates for all parameters in the model. Moreover, Bayesian item response theory (IRT) has benefits over frequentist approaches to IRT, including estimation for moderate and smaller sample sizes (Fox, 2010) and improved estimation of parameters (Natesan et al., 2016). van de Schoot et al. (2014) and Oldehinkel (2016) provide accessible discussions of Bayesian approaches in developmental science.

As a criterion-related test of validity of our approach to developmental scaling, we examine children's trajectories in relation to adjustment outcomes, including school readiness (math and reading skills) and externalizing problems. We examine self-regulation development across ages 3–7 years because self-regulatory processes show rapid development in early childhood (Greene, 2017; Montroy et al., 2016). Moreover, the transition to formal education may represent a key developmental period when self-regulatory processes become especially important for school readiness as well as for future learning and achievement (Blair & Raver, 2015; Mazzocco & Kover, 2007). Consistent with prior studies of self-regulation, we hypothesized that children's growth trajectories would show rapid development around age 3 and decelerate around age 7. We expected that measures would change in their strength of association with the self-regulation construct over time, consistent with heterotypic continuity (Petersen et al., 2016; Petersen, LeBeau, et al., 2021). Additionally, we hypothesized that boys would show poorer self-regulation, on average, than girls (Kochanska et al., 2001; Matthews et al., 2009, 2014; McClelland

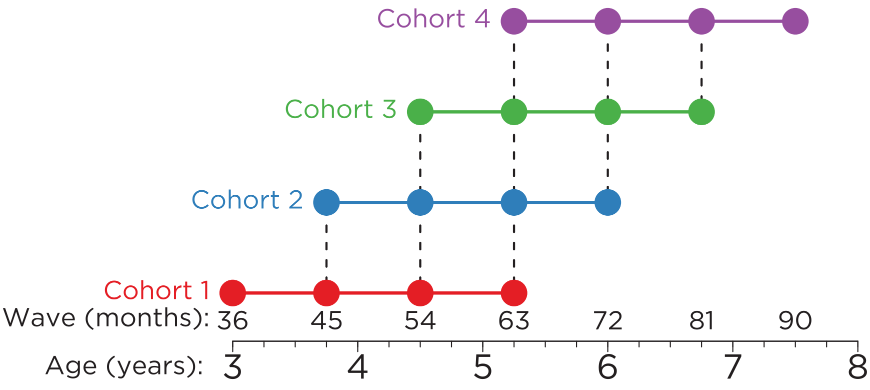


FIGURE 1 Accelerated longitudinal research design. *Note*. Accelerated longitudinal research design with four cohorts. The longitudinal design follows any given child for 2¼ years, with testing every 9 months; the whole data set spans the ages of 3–7½ years. Circles reflect measurement points (four waves) for each cohort. Dashed lines indicate common measurement points across cohorts.

et al., 2007). We also hypothesized that children's developmentally scaled self-regulation would be associated with school readiness and externalizing problems. Specifically, we hypothesized that lower levels of self-regulation would be associated with poorer math and reading skills, as well as externalizing problems.

2 | METHOD

2.1 | Participants

Participants consisted of a community sample of young children (*N* = 108) and their families, who took part in an ongoing accelerated longitudinal study. Children were recruited from 2018 to 2022 at one of the following ages (cohorts): 36 (*n* = 29), 45 (*n* = 29), 54 (*n* = 21), or 63 (*n* = 29) months and were assessed every 9 months over four time points (see Figure 1). The full sample of children spanned 3–7.5 years of age. Participants were recruited from Iowa City, Iowa and surrounding areas. Participants were recruited through a biomedical registry of children who had well-child checkups at University of Iowa Hospital and Clinics, university email listservs, and from advertisements and in-person recruitment activities at their school or preschool, Women, Infants, and Children (WIC) programs, pediatricians' offices, and community events. Exclusion criteria were: the child's primary caregiver did not speak English, or the child did not have a permanent guardian, did not have normal or corrected-to-normal vision and hearing, or was not capable of following basic instructions in English.

Figure 2 depicts the flow of participants from screening to consent. The final sample consisted of 108 children (*M* = 4.82 years, *SD* = 1.22 years; 51 girls), their primary caregiver, the primary caregiver's parenting partner (as applicable), and a teacher/secondary caregiver (e.g., nanny, babysitter, or someone else who knew the child well). Participant demographics are detailed in Appendix S1. The ethnic composition of children in the sample was: 67.6% Non-Hispanic White, 7.4% Black or African American, 6.5% Asian, 7.4% Hispanic or Latino, 5.6% Multiracial, and 5.6% other. Participants received money and small gift bags as compensation for participation.

Extent of missingness for each model variable is in Table S1. Reasons for missingness and tests of systematic missingness are in Appendix S2. The number of children with self-regulation scores by wave is depicted in Figure S1.

2.2 | Procedure

At each time point (i.e., every 9 months for four time points), the child and their primary caregiver completed two lab visits, approximately 1 week apart. The primary caregiver completed electronic questionnaires during both lab visits or from home. Additionally, the primary caregiver's parenting partner and the child's teacher/secondary caregiver were emailed or mailed the questionnaires to complete.

2.2.1 | Lab visit 1

The first lab visit lasted approximately 120–180 min (*M* = 150.78, *SD* = 20.31). During this visit, the child and their primary caregiver came to the lab. The child completed a series of tasks with an experimenter, including self-regulation tasks, parent–child interaction tasks, standardized assessments of academic achievement, and other assessments, while the primary caregiver completed questionnaires about their child.

2.2.2 | Lab visit 2

The second lab visit lasted approximately 70–120 min (*M* = 86.96, *SD* = 18.97). During this visit, the child completed computerized tasks, including a go/no-go (Fish/Sharks) and stop-signal (Food Finder) task, while wearing an electroencephalography cap and brainwaves were recorded. The primary caregiver completed additional questionnaires.

2.3 | Measures

The present study is part of a larger study, the School Readiness Study. Measures and hypotheses for the School Readiness Study were pre-registered: https://osf.io/jzxb8. Data files, a data dictionary, analysis scripts, and a computational notebook for the present study are

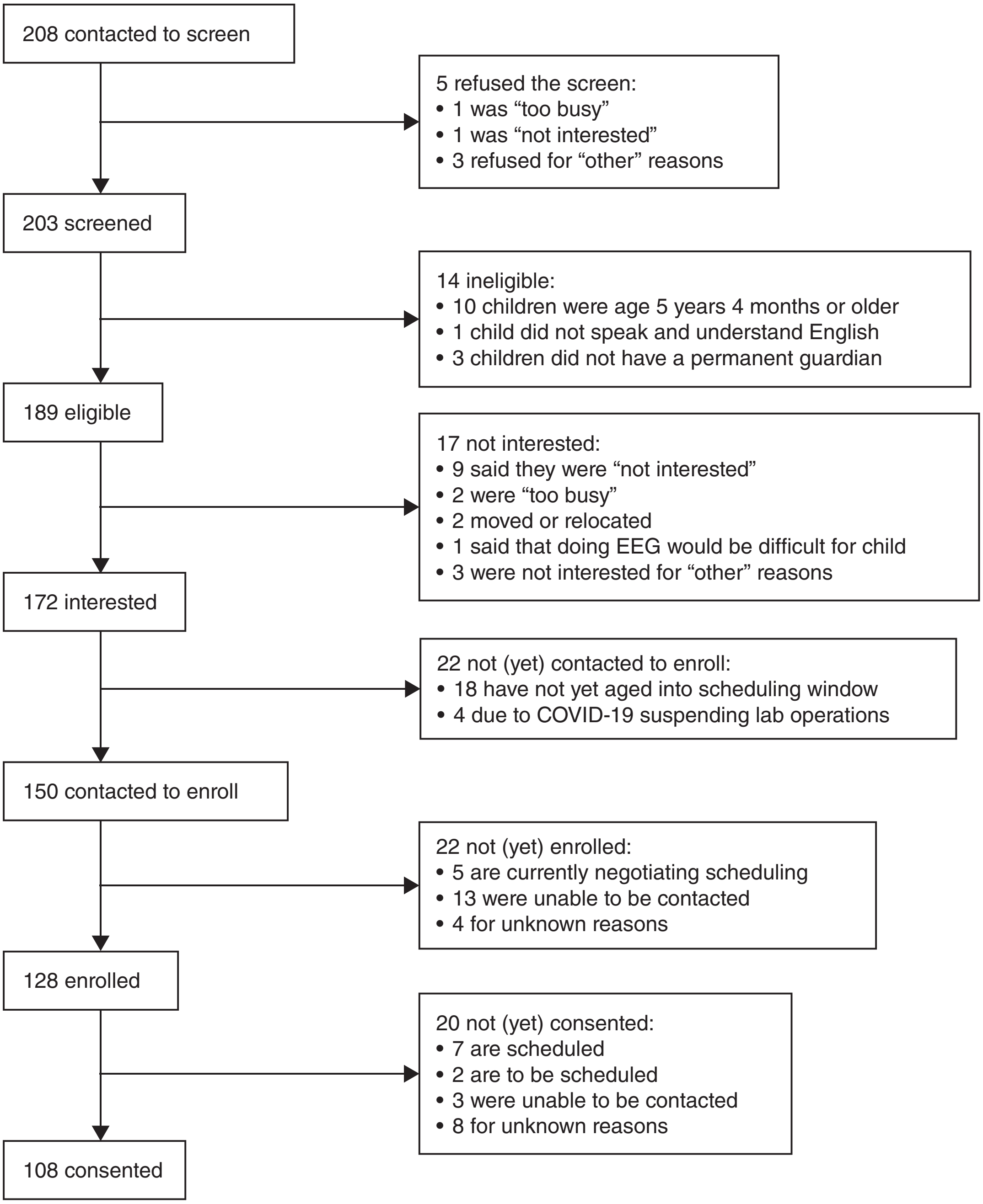

208 contacted to screen
5 refused the screen:
• 1 was "too busy"
• 1 was "not interested"
• 3 refused for "other" reasons
203 screened
14 ineligible:
• 10 children were age 5 years 4 months or older
• 1 child did not speak and understand English
• 3 children did not have a permanent guardian
189 eligible
17 not interested:
• 9 said they were "not interested"
• 2 were "too busy"
• 2 moved or relocated
• 1 said that doing EEG would be difficult for child
• 3 were not interested for "other" reasons
172 interested
22 not (yet) contacted to enroll:
• 18 have not yet aged into scheduling window
• 4 due to COVID-19 suspending lab operations
150 contacted to enroll
22 not (yet) enrolled:
• 5 are currently negotiating scheduling
• 13 were unable to be contacted
• 4 for unknown reasons
128 enrolled
20 not (yet) consented:
• 7 are scheduled
• 2 are to be scheduled
• 3 were unable to be contacted
• 8 for unknown reasons
108 consented

FIGURE 2 Participant flow chart. *Note*. EEG = "electroencephalography"

published online: https://osf.io/5xnrh. Descriptive statistics of model variables are in Tables S2–S4. Full descriptions of measures and covariates are in Appendix S3.

2.3.1 | Self-regulation

Measures of self-regulation included 15 laboratory tasks and two questionnaires. We assessed four facets of self-regulation: inhibitory control, delayed gratification, sustained attention, and executive functions. We assessed inhibitory control with Bear/Dragon, Day/Night, Fish/Sharks, Food Finder Stop-Signal Task, Grass/Snow, Hand Game, Knock/Tap, Less is More, Peg Tapping, Shape Stroop, Simon Says, and the Children's Behavior Questionnaire. We assessed delayed gratification with Gift Delay, a self-imposed waiting task, and Snack Delay. We assessed sustained attention with Token/Bead Sort. We assessed various executive functions, such as inhibition, shifting, and working memory, by parents' reports on the Behavior Rating Inventory of Executive Function (BRIEF). The BRIEF is a widely used questionnaire that was designed to assess executive functions within the context of children's everyday environment. Except for computer-scored tasks (Fish/Sharks and Stop-Signal Task) and Token/Bead Sort, children's performance on tasks was scored after the lab visit from video recording. All scored cases were double-coded to evaluate inter-rater reliability via intraclass correlation. Raters met to resolve any large discrepancies between raters' codes. Estimates of reliability (inter-rater, internal consistency, cross-time stability) are in Table S5. Scores were averaged across raters. Estimates of time to administer each task are in Table S6.

For developmental scaling, scores of each self-regulation measure were converted to proportion of maximum (POM) scores to have the same possible range (0–1), with higher scores reflecting greater self-regulation. Proportion scores are widely recommended by longitudinal researchers for studying growth with different measures (Little, 2013; Moeller, 2015). For measures that had a minimum and maximum possible score, the POM score reflected the proportion of the maximum possible score. For measures that did not have a minimum or maximum possible score (Stop-Signal Task and Token/Bead Sort), the POM score reflected the proportion of the maximum *observed* score. POM scores were calculated as: $\frac{\text{score} - \text{minimum}}{\text{maximum} - \text{minimum}}$, where minimum and maximum were the minimum and maximum possible or observed score. Tasks (Token/Bead Sort; Stop-Signal Task) and questionnaires (BRIEF) were adapted to accommodate the developmental capacity of the child and the changing expression of self-regulation with age.

2.3.2 | School Readiness

Woodcock Johnson IV–Tests of Achievement

The Woodcock Johnson IV–Tests of Achievement (Schrank et al., 2014, 2018) assess academic achievement. Children completed two subtests to assess their early (pre-)reading and math skills: Letter-Word Identification and Applied Problems, respectively. Letter-Word Identification assesses word identification skills and reading-writing ability. The child was asked to identify letters and eventually asked to read aloud individual words. Applied Problems assesses quantitative ability. The child was asked to analyze and solve applied math problems. Items were scored on accuracy (1 = correct, 0 = incorrect). Raw scores (i.e., number of correct responses) were used.

2.3.3 | Externalizing behavior

Achenbach System of Empirically Based Assessment

The Achenbach System of Empirically Based Assessment (ASEBA) assesses children's emotional and behavioral problems. Items were rated on a 3-point Likert scale according to how well the item described the child (0 = not true, 1 = somewhat or sometimes true, 2 = very true). Multiple versions were used based on the child's age and rater type. Parents completed the Child Behavior Checklist 1.5–5 (Achenbach & Rescorla, 2000) if the child was 3–5 years old or the Child Behavior Checklist 6–18 (Achenbach & Rescorla, 2000) if the child was 6–7 years old. Secondary caregivers completed the Caregiver–Teacher Report Form (Achenbach & Rescorla, 2001) if the child was 3–5 years old or the Teacher's Report Form (Achenbach & Rescorla, 2001) if the child was 6–7 years old. Scores on the Externalizing scale were used. Externalizing problem scores were then converted to POM scores to put scores from different ASEBA measures onto a metric with the same possible range.

2.4 | Statistical analysis

We used different measures of self-regulation across ages to account for heterotypic continuity.

2.4.1 | Exploratory factor analysis

We first examined whether measures' scores were able to be modeled with item response modeling by examining their scores in exploratory factor analysis (EFA). We conducted EFA with maximum likelihood estimation using the psych 2.1.9 package (Revelle, 2020) in R 4.1.2 (R Core Team, 2021).

2.4.2 | Developmental scaling

We used developmental scaling to link scores from the different measures across ages onto the same scale. In this way, we could make meaningful comparisons of scores from different measures across ages and estimate accurate trajectories of children's self-regulation growth. A detailed description of the developmental scaling approach is in Appendix S4. To perform developmental scaling, we used a two-parameter Bayesian longitudinal item response model in a mixed modeling item response theory (IRT) framework. Such a model allows us to simultaneously account for heterotypic continuity of self-regulation using different measures across time and to model children's self-regulation trajectories. Given the numerous measures assessed, the many items, and the varying number of items per measure, we used measure-level (POM) scores (rather than item- and trial-level scores) as the "items" in the item-response model. The model linked scores from measures across all ages in the same model, known as concurrent calibration. Concurrent calibration accounts for within-person dependence of scores across time and results in more precise and stable estimates than two-stage calibration in which separate models are fit (Kolen & Brennan, 2014; McArdle et al., 2009). The two-parameter item response model estimates two parameters: easiness (ξ; the inverse of difficulty) and discrimination (α). The item's easiness parameter is the expected score on an item at a given level of the construct (Bürkner, 2020). The item's discrimination parameter is how strongly the item is associated with the construct. In our study, easiness and discrimination provide information about the functioning and usefulness of each measure—and the whole measurement scheme—at a given age.

In the present study, the self-regulation scores were continuous proportion scores that ranged from 0 to 1. Because some scores were zero or one (especially one; see Figure S2), we used a zero-one-inflated beta distribution for the outcome variable (Ospina & Ferrari, 2012). A traditional beta distribution is a continuous probability distribution that does not allow zeros or ones. A zero-one-inflated beta distribution is a mixed continuous-discrete probability distribution, which includes a continuous beta distribution (to capture the continuous distribution of proportion scores) and a Bernoulli distribution (to capture zeros and ones).

We performed the developmental scaling, estimation of growth curves, and tests of differential item (measure) functioning (DIF) in the same model. A given child had up to four time points. Thus, a quadratic was the most complex polynomial of nonlinear growth we could estimate for children's trajectories that still allow measurement error. Because of prior work demonstrating that growth in self-regulation is non-linear, such that children showed faster growth in preschool than elementary school (Montroy et al., 2016), we modeled children's growth in self-regulation with a quadratic term. We modeled random intercepts and random linear and quadratic slopes to allow each child to differ in their starting point, form of growth, and curvature. Age in years was centered to set the intercepts at age 3. We included the child's sex (female = 1, male = 0) as a predictor of the intercepts and slopes.

To examine DIF across ages, we estimated a random intercept and slope for measure (in addition to the terms described above) to allow the item parameters for each measure to differ by age. This allowed us to examine the extent to which the measures changed across development in easiness and discrimination.

Our model had no missing data in the predictors (age and sex); missingness was only in the outcome (scores on self-regulation measures). Mixed models handle missing data in the outcomes. Mixed models provide valid inferences if the data are missing at random or completely at random. Because much of our missingness was due to COVID-19, and we observed limited patterns of systematic missingness as a function of demographics, predictors, or outcomes with small effect sizes, we felt this modeling approach was appropriate. Moreover, researchers have argued against using multiple imputation in longitudinal designs that use mixed models because multiple imputation can lead to unstable estimates (Twisk et al., 2013).

Developmentally scaled self-regulation factor scores were estimated for each child at each of their measurement occasions. This allowed each child to have a different factor score at each of their measurement occasions.

We fit the Bayesian longitudinal mixed model using the brms package 2.16.3 (Bürkner, 2017) in R, which uses the RStan 2.21.3 (Stan Development Team, 2020a) interface to Stan 2.21.0 (Stan Development Team, 2020b) for Bayesian modeling. The model included eight chains and 10,000 iterations.

2.4.3 | Sensitivity analyses

We conducted sensitivity analyses using two models, as described in Appendix S5. First, we examined a model that imposed approximate longitudinal measurement invariance. Second, we fit a model that excluded scores for a given measure at ages when the proportion of maximum score on that measure could reflect ceiling effects (i.e., mean proportion score > 0.90).

2.4.4 | Self-regulation predicting outcomes

To examine whether children's developmentally scaled self-regulation factor scores were associated with externalizing problems and school readiness, we used multiple regression with a cluster variable specifying the participant (i.e., clustered regression). Clustered regression accounts for the longitudinal dependency in the data. Clustered regression models were fit using the rms package 6.2 (Harrell, 2015) in R that calculates robust standard errors using a Huber-White sandwich estimator of the covariance matrix (Huber, 1967; White, 1980). Power analyses of our ability to detect associations predicting outcomes are in Appendix S6.

3 | RESULTS

3.1 | Descriptive statistics and correlations

Average proportion self-regulation scores by measure and age are shown in Figure 3. Bivariate correlations and descriptive statistics of model variables are in Tables S2–S4. Partial correlations controlling for age are in Tables S7–S9. Although there were exceptions, self-regulation scores were largely inter-correlated across measures. Moreover, a combination of self-regulation scores across measures showed strong internal consistency ($\omega = 0.94$; see Table S5).

3.2 | Exploratory factor analysis

We examined scores from the self-regulation measures in EFA. Results of the EFA are in Table S10. A one-factor model accounted for 35% of the variance. All but three measures' scores (Food Finder Stop-Signal Task, mothers' and fathers' ratings on the BRIEF, and mothers' and secondary caregivers' ratings on the CBQ) had a standardized factor loading above 0.40. In a two-factor model, the second factor accounted for 8% of the variance. Moreover, all measures that had loadings above 0.40 on the second factor were questionnaire measures, suggesting that the factor that accounted for the most variance after the primary factor was a method factor. Findings remained consistent when controlling for the child's age. Thus, although the self-regulation measures clearly assessed multiple dimensions, a single factor accounted for considerable variance, and accounted for considerably more variance than the second factor. Based on this evidence, the primary factor appeared to reflect a meaningful operationalization of self-regulation. Given our goals to examine children's self-regulation development by aggregating scores from multiple methods, we conducted item response modeling with a single factor.

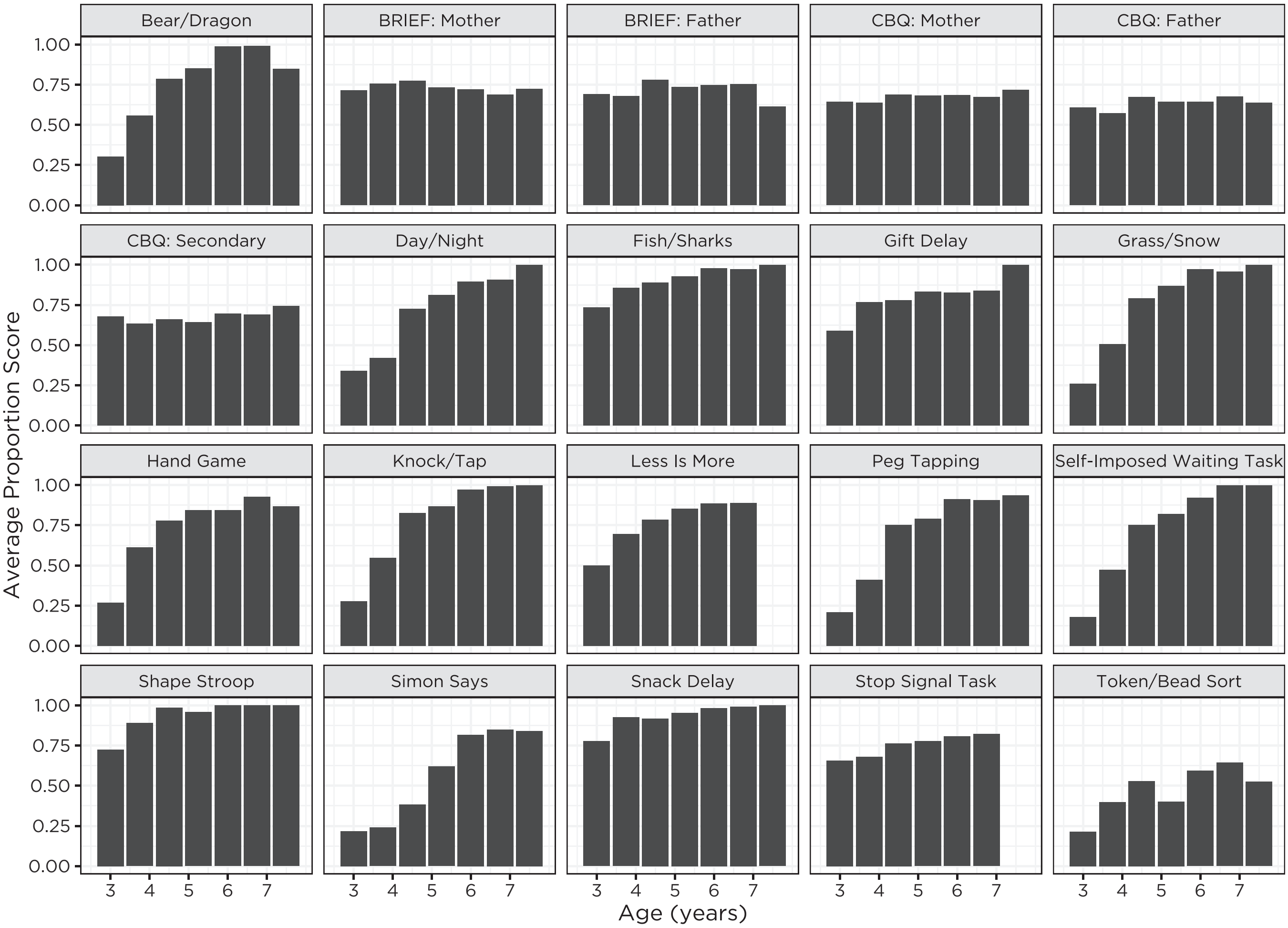


FIGURE 3 Average proportion self-regulation scores by measure and age. *Note*. The bars correspond to waves 36, 45, 54, 63, 72, 81, and 90 months. Scores on some performance-based measures were largely missing at 90 months due to COVID-19 (see Appendix S1). "BRIEF" = Behavior Rating Inventory of Executive Function; "CBQ" = Children's Behavior Questionnaire; "Secondary" = secondary caregiver.

3.3 | Bayesian longitudinal item response model

We fit a Bayesian longitudinal item response model in a mixed modeling framework to perform the developmental scaling, estimation of growth curves, and tests of DIF. All Gelman-Rubin diagnostic criteria for convergence ($\hat{R}$) were 1.00, and visual examination of trace plots showed that all chains adequately mixed, indicating that the model converged. The R^2 from leave-one-out (LOO) cross validation was 0.47, indicating that the model explained nearly half of the variance in children's scores on the self-regulation measures across time. Measures' easiness and discrimination are in Figure 4. All measures showed significant associations with the self-regulation construct; the 95% credible interval of the discrimination estimates did not include zero. Measures' empirical characteristic curves are in Figure S4. Model results are in Table S11.

3.3.1 | Differential item (measure) functioning

Tests of differential item functioning are described in detail in Appendix S5. Changes in item easiness and discrimination are depicted in Figure S3. Four measures became easier—relative to the same ability—with age: Fish/Sharks, Gift Delay, Snack Delay, and mothers' ratings on the BRIEF–P. All measures except Fish/Sharks and Simon Says showed decreases in discrimination with age, consistent with heterotypic continuity. Effect sizes of non-invariance were small, and measures remained strongly discriminating across ages, so we proceeded to interpret the growth curves and predictors of the trajectories.

3.3.2 | Form of growth

There was a positive mean of the quadratic slope. As depicted in Figure 5, children showed rapid growth in self-regulation from ages 3 to 6, after which growth slowed and leveled off.

3.3.3 | Sex-related differences

We examined whether the child's sex predicted differences in intercepts and slopes. As depicted in Figure 5, girls showed higher intercepts

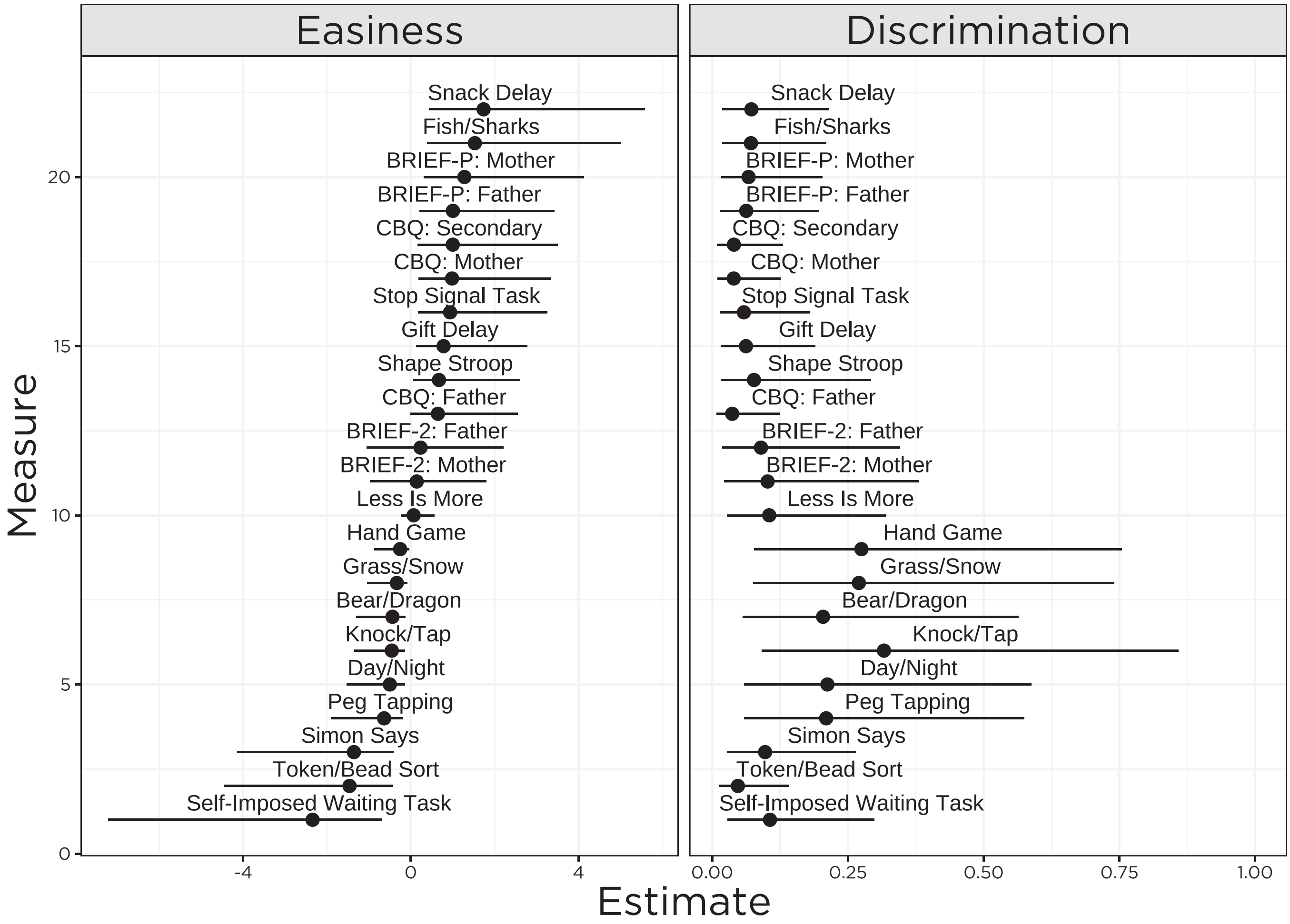


FIGURE 4 Easiness and discrimination of the self-regulation measures at age 3. *Note*. The lines represent the 95% credible interval. Note that the metric of easiness and discrimination in the zero-one-inflated beta item response model is different from the metric of difficulty and discrimination in the traditional two-parameter logistic item response theory models. "BRIEF" = Behavior Rating Inventory of Executive Function; "CBQ" = Children's Behavior Questionnaire; "Secondary" = secondary caregiver.

of self-regulation than boys at age 3. The effect size was small (a difference of 3.8%), and boys appeared to nearly catch up to girls by age 7. Girls and boys did not significantly differ in their linear or quadratic slopes.

3.4 | Validation of developmentally scaled self-regulation scores

As a validation of the developmentally scaled self-regulation scores, we examined whether the developmentally scaled self-regulation scores were associated with theoretically relevant outcomes, including externalizing problems and school readiness.

3.4.1 | Predicting externalizing problems

Model results from the regression models of developmentally scaled self-regulation scores predicting externalizing problems are in Table S12–S13. Self-regulation was moderately negatively associated with externalizing problems in a model without covariates (β = -0.28). However, the association became marginally significant when controlling for the child's age (β = -0.13).

3.4.2 | Predicting school readiness

Model results from the regression models of developmentally scaled self-regulation scores predicting school readiness are in Table S14–S16. Self-regulation was moderately to strongly positively associated with reading (β = 0.27) and math (β = 0.51) skills, controlling for age, grade, and SES. Moreover, self-regulation remained associated with math skills (β = 0.35) but was marginally associated with reading skills (β = 0.19), when controlling for intelligence.

3.4.3 | Sensitivity analyses

The sensitivity analyses are described in detail in Appendix S5. Findings were substantially similar when examining the model with approximate

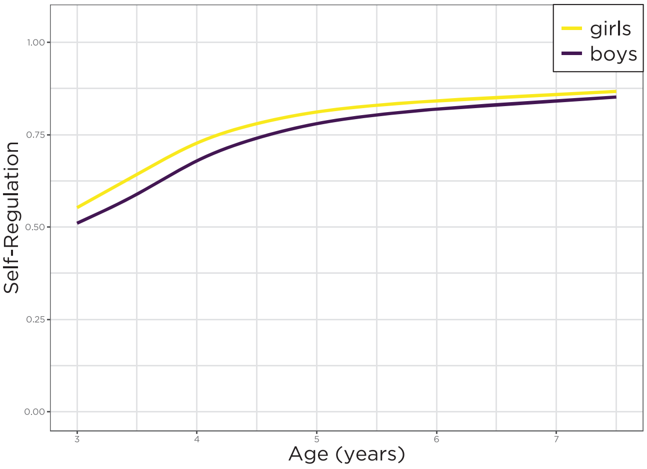


FIGURE 5 Model-implied self-regulation growth curves by the child's sex

longitudinal invariance imposed, and when examining the model with potential mean-level ceiling effects. Both models yielded similar trajectories of self-regulation. In addition, criterion-related tests yielded similar results. Self-regulation was negatively associated with externalizing problems, controlling for age, and was positively associated with math (but not reading) skills when controlling for age, grade, SES, and intelligence.

4 | DISCUSSION

Self-regulation is thought to demonstrate changes in its behavioral manifestation across development. Researchers have argued that children develop lower-level processes in early childhood (e.g., early forms of inhibitory control and delayed gratification) and higher-level processes (e.g., executive functions) in later childhood, which are then integrated with lower-level processes to form a hierarchically organized regulatory system, in later childhood (Greene, 2017; Montroy et al., 2016). However, limited empirical work has examined whether self-regulation shows heterotypic continuity, despite evidence for heterotypic continuity of specific components of self-regulation (Chang et al., 2015; Geeraerts et al., 2021; Petersen, Bates, et al., 2021; Petersen et al., 2016; Putnam et al., 2008; Zimmermann & Iwanski, 2014). We found evidence of heterotypic continuity of self-regulation in the present study, such that measures changed in their strength of association with the latent construct across ages. Prior work has not accounted for heterotypic continuity of self-regulation when studying children's growth curves. In the present study, we accounted for heterotypic continuity of self-regulation based on theoretical, methodological, and analytical considerations. We followed theoretical conceptualizations of self-regulation as encompassing multiple processes, including inhibitory control, delayed gratification, sustained attention, and executive functions (Baumeister & Vohs, 2004; Berger, 2011; Blair & Raver, 2015; Gagne et al., 2021; McClelland et al., 2010, 2015; Nigg, 2017). Methodologically, we used different measures across ages to account for the changing nature of the construct, while using some common measures across adjacent ages to ensure scores could be linked across ages. Analytically, we used developmental scaling to link scores from different measures across ages onto the same scale so we could examine children's self-regulation growth. We describe our approach to developmental scaling below.

4.1 | Developmental scaling

We used a Bayesian longitudinal item response modeling approach to developmental scaling, in which children's proportion scores from each measure were used as items in the model. The model simultaneously estimated item response model parameters and longitudinal growth curves using a concurrent calibration approach, in which scores from measures across all ages were linked in the same model (Kolen & Brennan, 2014). Our approach to developmental scaling is consistent with prior work that has linked different measures of cognitive ability across the lifespan (McArdle et al., 2009). When concurrent calibration is used, people's estimated construct levels are on the same scale across ages if IRT assumptions are met (Kolen & Brennan, 2014). Although the measures likely assessed multiple dimensions, EFA demonstrated that the data were likely uni-dimensional enough for IRT, thus providing greater confidence in children's estimated growth curves. Moreover, the developmentally scaled scores showed strong internal consistency ($\omega = 0.94$) and cross-time stability ($r = 0.68$), providing evidence that they reflected meaningful operationalization of children's self-regulation.

4.2 | Form of growth

Based on model-implied trajectories from the longitudinal item response model, children showed rapid growth in self-regulation from ages 3 to 6, after which growth slowed and leveled off from ages 6 to 7. This pattern of growth is consistent with concomitant changes in brain development. Brain size increases four-fold during the preschool period, reaching approximately 90% of the adult volume by age 6 (Brown & Jernigan, 2012; Stiles & Jernigan, 2010). Moreover, children between ages 3 and 6 show marked improvement in working memory and inhibitory control abilities, which are thought to depend on the dorsolateral prefrontal cortex (DL-PFC; Berger, 2011; Diamond, 2001, 2002). During this time, the DL-PFC undergoes important changes, including rapid decreases in neuronal density between 2 and 7 years of age and expansion of dendritic trees in layer III pyramidal cells between 2 and 5 years of age (Diamond, 2001).

Consistent with prior work, girls had modestly higher mean self-regulation at age 3 compared to boys (Kochanska et al., 2001; Matthews et al., 2009, 2014; McClelland et al., 2007). Boys and girls did not significantly differ in their slopes, but boys appeared to nearly catch up to girls by age 7.

4.3 | Validation of developmentally scaled self-regulation scores

As a criterion-related test of the validity of our approach to developmental scaling, we examined children's developmentally scaled self-regulation scores in relation to adjustment outcomes, including externalizing problems and school readiness (math and reading skills). We hypothesized that lower levels of self-regulation would be associated with externalizing problems and poorer math and reading skills. We found that lower levels of self-regulation were moderately associated with externalizing problems. However, the association was only at trend level when controlling for the child's age, which was inconsistent with hypotheses. It is possible that developmental improvements in self-regulation could account for normative age-related reductions in externalizing problems. Alternatively, developmental changes in self-regulation may reflect other processes, such as language development (Petersen & LeBeau, 2021), that lead to age-related reductions in externalizing problems. Or, perhaps our study was under-powered to detect the association, given meta-analytic evidence that the effect size of self-regulation on externalizing problems is small (Berger & Buttelmann, 2021). Future work will be important to examine the role of self-regulation in the development of externalizing problems.

Consistent with hypotheses, lower levels of self-regulation were associated with poorer (pre-)reading and math skills. The effect size was large, and the association held when controlling for covariates (age, grade, and SES). Moreover, the association held with math (but not reading) skills when controlling for the child's intelligence. Thus, performance on the self-regulation measures does not appear to merely reflect better comprehension of task rules (likely influenced by language ability, i.e., a dimension of intelligence). The finding that self-regulation was strongly associated with math skills above and beyond intelligence supports the possibility raised by prior research that self-regulation plays an important role in development of school readiness (e.g., Blair & Raver, 2015; Eisenberg et al., 2010; Ursache et al., 2012). The finding that self-regulation was more strongly associated with academic skills than with externalizing problems is consistent with prior research showing that preschoolers' executive function predicted math skills but not aggression (Sasser et al., 2015). In sum, the criterion-related association between children's developmentally scaled self-regulation scores and their school readiness provides further support for the validity and utility of our approach to developmental scaling.

4.4 | Implications for understanding development of self-regulation

Theory (Berger, 2011; Kopp, 1982; McClelland et al., 2010), empirical work (Chang et al., 2015; Geeraerts et al., 2021; Petersen, Bates, et al., 2021; Putnam et al., 2008; Zimmermann & Iwanski, 2014), meta-analysis (Petersen et al., 2016), and findings in the present study collectively provide considerable evidence that self-regulation changes in its behavioral manifestation across development. Our developmental scaling approach of self-regulation replicates and extends prior literature examining development of self-regulation. Consistent with previous studies, we observed rapid growth in self-regulation between ages 3 and 6, which slowed and leveled off between ages 6 and 7 (Greene, 2017; Montroy et al., 2016). Moreover, we observed robust associations with school readiness outcomes. Crucially, developmental scaling simultaneously accounted for heterotypic continuity and charted children's growth over time. Accounting for changing behavioral manifestations at different ages provides a more accurate understanding of self-regulation development across early childhood than previous models, and our approach can be extended to adolescence and adulthood. Indeed, research has shown that adolescents show a marked increase in cognitive flexibility and improvements in planning, organizing, and strategic thinking skills, which carry into adulthood (Anderson, 2002; Greene, 2017). Moreover, Zimmermann and Iwanski (2014) found differences in emotion-regulation strategies from early adolescence to middle adulthood, consistent with heterotypic continuity. Nevertheless, self-regulation or other constructs do not need to show heterotypic continuity for our modeling approach to be useful for charting children's growth.

4.5 | Strengths

The study had several strengths. First, the study was longitudinal, which allowed examining children's self-regulation development. Second, we assessed multiple facets of self-regulation to be consistent with theory and prior research on the structure of self-regulation. Third, our assessment of self-regulation included multiple measurement methods including performance-based assessment and questionnaires to reduce common method variance. Fourth, we included multiple informants, including mothers, fathers, and teachers or other caregivers to gain a more accurate estimate of children's real-world functioning. Fifth, we used developmental scaling to link differing measures across ages onto the same scale, which allowed examining children's absolute growth in self-regulation. Our multi-wave, multi-facet, multi-method, multi-measure, multi-rater, developmental scaling approach is the most comprehensive to date for assessing development of self-regulation. Prior research using developmental scaling has used primarily dichotomous or polytomous items. The Bayesian approach we used successfully handled a moderate sample size when fitting longitudinal item response models with continuous data, which potentially increases its practicality for use in developmental research. We also make our data and analysis scripts freely available to promote dissemination.

4.6 | Limitations

The study also had limitations. First, the sample size may limit our ability to detect smaller effects. Second, the study was observational, so we cannot make causal inferences. Third, there was considerable missing data at later ages, including limited performance-based assessments at participants' fourth time points, largely due to COVID-19. In addition, a modest number of children had self-regulation scores at later ages due to the accelerated nature of the longitudinal design. Moreover, many measures showed increases in easiness and/or decreases in discrimination across ages, and several performance-based tasks showed ceiling effects at later ages, which may have contributed to their somewhat weaker discrimination. Thus, we have less confidence about children's level of self-regulation at later ages in our study (6–7 years of age). Nevertheless, researchers have argued that establishing longitudinal measurement invariance is unnecessary when the construct shows heterotypic continuity (Edwards & Wirth, 2012; Knight & Zerr, 2010; Petersen et al., 2020). Our model accounted for changes in measures' easiness and discrimination. Moreover, the form of growth we observed aligns with prior findings, and it was consistent even when we imposed approximate longitudinal measurement invariance and removed scores at ages with potential mean-level ceiling effects, which increases confidence in our findings.

Another limitation relates to assumptions regarding self-regulation. We modeled self-regulation using item response modeling, which assumes there is a latent factor (i.e., reflective construct) that influences scores on all self-regulation measures. In support of a reflective model of self-regulation, we found that the measures assessing various components of self-regulation (i.e., inhibitory control, delayed gratification, sustained attention, and executive functions) were robustly correlated, and a one-factor model captured a large portion of the variance in scores across measures. We acknowledge, however, that we may not have assessed all relevant components, for instance emotion regulation, which could limit the interpretation of the latent factor and generalizability of findings. Nevertheless, our assessment included many measures of multiple facets, providing a more comprehensive assessment than prior research examining self-regulation growth, which has mainly examined one or a few measures and one or a few facets (Montroy et al., 2016; Sulik et al., 2010). Alternatively, emerging research suggests that self-regulatory processes may be operationalized using formative constructs (Camerota et al., 2020; Willoughby et al., 2017), whereby self-regulation is defined as the summation of relevant measures, rather than as their shared variance. Future research should examine how to operationalize self-regulation, including its structure, whether it is a reflective or formative construct, and how its structure changes with development. Better developmental models of self-regulation that account for changes in its structure will lead to better understanding of how self-regulation develops across the lifespan.

5 | CONCLUSION

Self-regulation is thought to change in its behavioral manifestation across development. We accounted for heterotypic continuity of self-regulation by using different, theoretically relevant measures across ages to account for the changing manifestation of the construct. We used developmental scaling to link scores from differing measures across ages onto the same scale so we could examine children's self-regulation growth. Children's developmentally scaled self-regulation scores were validated against their externalizing problems and school readiness, including math and reading skills. Findings suggest that developmental scaling permits studying the development of self-regulation across lengthy spans and key developmental transitions. Future research should adapt measurement schemes to be developmentally appropriate and valid across ages. Developmental scaling may enable studying development of self-regulation and other constructs across the lifespan.

ACKNOWLEDGMENTS

The present study was approved by the University of Iowa Institutional Review Board (Study #: 201708761). The project was funded by Grants HD098235 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), T32GM108540 from the National Institute of General Medical Sciences (NIGMS), and UL1TR002537 from the National Center for Advancing Translational Sciences (NCATS).

CONFLICT OF INTEREST

We have no conflicts of interest to disclose.

DATA AVAILABILITY STATEMENT

Data files, a data dictionary, analysis scripts, and a computational notebook for the present study are published online: https://osf.io/5xnrh

ORCID

Alexis Hosch https://orcid.org/0000-0003-2874-2340
Jacob J. Oleson https://orcid.org/0000-0001-6343-3274
Jordan L. Harris https://orcid.org/0000-0002-1335-9448
Isaac T. Petersen https://orcid.org/0000-0003-3072-6673

REFERENCES

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA preschool forms and profiles: An integrated system of multi-informant assessment*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. Burlington, VT: University of Vermont, Department of Psychiatry.

Allan, N. P., & Lonigan, C. J. (2011). Examining the dimensionality of effortful control in preschool children and its relation to academic and socio-emotional indicators. *Developmental Psychology*, *47*(4), 905–915. https://doi.org/10.1037/a0023748

Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. *Child Neuropsychology*, *8*(2), 71–82. https://doi.org/10.1076/chin.8.2.71.8724

Backer-Grøndahl, A., Nærde, A., & Idsoe, T. (2019). Hot and cool self-regulation, academic competence, and maladjustment: Mediating and differential relations. *Child Development*, *90*(6), 2171–2188. https://doi.org/10.1111/cdev.13104

Bates, J. E., & Novosad, C. (2005). Measurement of individual difference constructs in child development, or taking aim at moving targets. In D. M. Teti (Ed.), *Handbook of research methods in developmental science* (pp. 103–122). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470756676.ch6

Baumeister, R. F., & Vohs, K. D. (Eds.). (2004). *Handbook of self-regulation: Research, theory, and applications*. Guilford Press.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1–3), 7–15. https://doi.org/10.1016/0010-0277(94)90018-3

Berger, A. (2011). *Self-regulation: Brain, cognition, and development*. American Psychological Association. https://doi.org/10.1037/12327-000

Berger, P., & Buttelmann, D. (2021). A meta-analytic approach to the association between inhibitory control and parent-reported behavioral adjustment in typically-developing children: Differentiating externalizing and internalizing behavior problems. *Developmental Science*, e13141. https://doi.org/10.1111/desc.13141

Berk, L. E. (1999). Children's private speech: An overview of theory and the status of research. In P. Llyod & C. Fernyhough (Eds.), *Lev Vygotsky: Critical assessments: Thought and language* (vol. 2, pp. 30–70). Taylor & Frances/Routledge.

Best, J. R., Miller, P. H., & Jones, L. L. (2009). Executive functions after age 5: Changes and correlates. *Developmental Review*, *29*(3), 180–200. https://doi.org/10.1016/j.dr.2009.05.002

Bivens, J. A., & Berk, L. E. (1990). A longitudinal study of the development of elementary school children's private speech. *Merrill-Palmer Quarterly*, *36*, 443–463.

Blair, C., & Diamond, A. (2008). Biological processes in prevention and intervention: The promotion of self-regulation as a means of preventing school failure. *Development and Psychopathology*, *20*(3), 899–911. https://doi.org/10.1017/S0954579408000436

Blair, C., & Raver, C. C. (2015). School readiness and self-regulation: A developmental psychobiological approach. *Annual Review of Psychology*, *66*, 711–731. https://doi.org/10.1146/annurev-psych-010814-015221

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*(2), 647–663. https://doi.org/10.1111/j.1467-8624.2007.01019.x

Bridgett, D. J., Burt, N. M., Edwards, E. S., & Deater-Deckard, K. (2015). Intergenerational transmission of self-regulation: A multidisciplinary review and integrative conceptual framework. *Psychological Bulletin*, *141*(3), 602–654. https://doi.org/10.1037/a0038662

Bridgett, D. J., Oddi, K. B., Laake, L. M., Murdock, K. W., & Bachmann, M. N. (2013). Integrating and differentiating aspects of self-regulation: Effortful control, executive functioning, and links to negative affectivity. *Emotion*, *13*(1), 47–63. https://doi.org/10.1037/a0029536

Brown, T. T., & Jernigan, T. L. (2012). Brain development during the preschool years. *Neuropsychology Review*, *22*(4), 313–333. https://doi.org/10.1007/s11065-012-9214-1

Bürkner, P.-C. (2017). brms: An R Package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2020). Bayesian item response modeling in R with brms and Stan. *ArXiv:1905.09501*. https://doi.org/10.48550/arXiv.1905.09501

Calkins, S. D., & Fox, N. A. (2002). Self-regulatory processes in early personality development: A multilevel approach to the study of childhood social withdrawal and aggression. *Development and Psychopathology*, *14*(3), 477–498. https://doi.org/10.1017/S095457940200305X

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

Camerota, M., Willoughby, M. T., Magnus, B. E., & Blair, C. B. (2020). Leveraging item accuracy and reaction time to improve measurement of child executive function ability. *Psychological Assessment*, *32*(12), 1118–1132. https://doi.org/10.1037/pas0000953

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. https://doi.org/10.1111/1467-8624.00333

Carlson, S. M., & Wang, T. S. (2007). Inhibitory control and emotion regulation in preschool children. *Cognitive Development*, *22*(4), 489–510. https://doi.org/10.1016/j.cogdev.2007.08.002

Chang, H., Shaw, D. S., & Cheong, J. (2015). The development of emotional and behavioral control in early childhood: Heterotypic continuity and relations to early school adjustment. *Journal of Child and Adolescent Behavior*, *3*(3), 204. https://doi.org/10.4172/2375-4494.1000204

Chen, F. R., & Jaffee, S. R. (2015). The heterogeneity in the development of homotypic and heterotypic antisocial behavior. *Journal of Developmental and Life-Course Criminology*, *1*(3), 269–288. https://doi.org/10.1007/s40865-015-0012-3

Cicchetti, D., & Rogosch, F. A. (2002). A developmental psychopathology perspective on adolescence. *Journal of Consulting and Clinical Psychology*, *70*(1), 6–20. https://doi.org/10.1037/0022-006X.70.1.6

Cole, P. M., Martin, S. E., & Dennis, T. A. (2004). Emotion regulation as a scientific construct: Methodological challenges and directions for child development research. *Child Development*, *75*(2), 317–333. https://doi.org/10.1111/j.1467-8624.2004.00673.x

Cole, P. M., Ram, N., & English, M. S. (2019). Toward a unifying model of self-regulation: A developmental approach. *Child Development Perspectives*, *13*(2), 91–96. https://doi.org/10.1111/cdep.12316

Denham, S. A., Warren-Khot, H. K., Bassett, H. H., Wyatt, T., & Perna, A. (2012). Factor structure of self-regulation in preschoolers: Testing models of a field-based assessment for predicting early school readiness. *Journal of Experimental Child Psychology*, *111*(3), 386–404. https://doi.org/10.1016/j.jecp.2011.10.002

Diamond, A. (2001). A model system for studying the role of dopamine in the prefrontal cortex during early development in humans: Early and continuously treated phenylketonuria. In C. A. Nelson & M. Luciana (Eds.), *Handbook of developmental cognitive neuroscience* (pp. 433–472). MIT Press.

Diamond, A. (2002). Normal development of prefrontal cortex from birth to young adulthood: Cognitive functions, anatomy, and biochemistry. In D. T. Stuss & R. T. Knight (Eds.), *Principles of frontal lobe function* (pp. 466–503). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195134971.003.0029

Edwards, M. C., & Wirth, R. J. (2012). Valid measurement without factorial invariance: A longitudinal example. In J. R. Harring & G. R. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences* (pp. 289–311). IAP Information Age Publishing.

Eisenberg, I. W., Bissett, P. G., Canning, J. R., Dallery, J., Enkavi, A. Z., Whitfield-Gabrieli, S., Gonzalez, O., Green, A. I., Greene, M. A., Kiernan, M., Kim, S. J., Li, J., Lowe, M. R., Mazza, G. L., Metcalf, S. A., Onken, L., Parikh, S. S., Peters, E., Prochaska, J. J., … Poldrack, R. A. (2018). Applying novel technologies and methods to inform the ontology of self-regulation. *Behaviour Research and Therapy*, *101*, 46–57. https://doi.org/10.1016/j.brat.2017.09.014

Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications*, *10*(1), 2319. https://doi.org/10.1038/s41467-019-10301-1

Eisenberg, N., Cumberland, A., Spinrad, T. L., Fabes, R. A., Shepard, S. A., Reiser, M., Murphy, B. C., Losoya, S. H., & Guthrie, I. K. (2001). The relations of regulation and emotionality to children's externalizing and internalizing problem behavior. *Child Development*, *72*(4), 1112–1134. https://doi.org/10.1111/1467-8624.00337

Eisenberg, N., Spinrad, T. L., & Eggum, N. D. (2010). Emotion-related self-regulation and its relation to children's maladjustment. *Annual Review of Clinical Psychology*, *6*, 495–525. https://doi.org/10.1146/annurev.clinpsy.121208.131208

Eisenberg, N., Valiente, C., Spinrad, T. L., Cumberland, A., Liew, J., Reiser, M., Zhou, Q., & Losoya, S. H. (2009). Longitudinal relations of children's effortful control, impulsivity, and negative emotionality to their externalizing, internalizing, and co-occurring behavior problems. *Developmental Psychology*, *45*(4), 988–1008. https://doi.org/10.1037/a0016213

Espy, K. A., Sheffield, T. D., Wiebe, S. A., Clark, C. A. C., & Moehr, M. (2011). Executive control and dimensions of problem behaviors in preschool children. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *52*(1), 33–46. https://doi.org/10.1111/j.1469-7610.2010.02265.x

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications* (1st ed.). Springer.

Gagne, J. R., Liew, J., & Nwadinobi, O. K. (2021). How does the broader construct of self-regulation relate to emotion regulation in young children?. *Developmental Review*, *60*, 100965. https://doi.org/10.1016/j.dr.2021.100965

Geeraerts, S. B., Endendijk, J. J., Deković, M., Huijding, J., Deater-Deckard, K., & Mesman, J. (2021). Inhibitory control across the preschool years: Developmental changes and associations with parenting. *Child Development*, *92*(1), 335–350. https://doi.org/10.1111/cdev.13426

Greene, J. A. (2017). *Self-regulation in education* (1st edition). Routledge.

Harrell, F. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). Springer International Publishing. https://doi.org/10.1007/978-3-319-19425-7

Hofmann, W., Schmeichel, B. J., & Baddeley, A. D. (2012). Executive functions and self-regulation. *Trends in Cognitive Sciences*, *16*(3), 174–180. https://doi.org/10.1016/j.tics.2012.01.006

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1: Statistics*, 221–233.

Knight, G. P., & Zerr, A. A. (2010). Informed theory and measurement equivalence in child development research. *Child Development Perspectives*, *4*(1), 25–30. https://doi.org/10.1111/j.1750-8606.2009.00112.x

Kochanska, G. (1997). Multiple pathways to conscience for children with different temperaments: From toddlerhood to age 5. *Developmental Psychology*, *33*(2), 228–240. https://doi.org/10.1037/0012-1649.33.2.228

Kochanska, G. (2002). Committed compliance, moral self, and internalization: A mediational model. *Developmental Psychology*, *38*(3), 339–351. https://doi.org/10.1037/0012-1649.38.3.339

Kochanska, G., Coy, K. C., & Murray, K. T. (2001). The development of self-regulation in the first four years of life. *Child Development*, *72*(4), 1091–1111. https://doi.org/10.1111/1467-8624.00336

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. Springer Science & Business Media.

Kopp, C. B. (1982). Antecedents of self-regulation: A developmental perspective. *Developmental Psychology*, *18*(2), 199–214. https://doi.org/10.1037/0012-1649.18.2.199

Lewis, M. D., & Stieben, J. (2004). Emotion regulation in the brain: Conceptual issues and directions for developmental research. *Child Development*, *75*(2), 371–376. https://doi.org/10.1111/j.1467-8624.2004.00680.x

Liew, J. (2012). Effortful control, executive functions, and education: Bringing self-regulatory and social-emotional competencies to the table. *Child Development Perspectives*, *6*(2), 105–111. https://doi.org/10.1111/j.1750-8606.2011.00196.x

Liew, J., Cameron, C. E., & Lockman, J. J. (2018). Parts of the whole: Motor and behavioral skills in self-regulation and schooling outcomes. *Early Education and Development*, *29*(7), 909–913. https://doi.org/10.1080/10409289.2018.1500513

Lin, B., Liew, J., & Perez, M. (2019). Measurement of self-regulation in early childhood: Relations between laboratory and performance-based measures of effortful control and executive functioning. *Early Childhood Research Quarterly*, *47*, 1–8. https://doi.org/10.1016/j.ecresq.2018.10.004

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

Martel, M. M., & Nigg, J. T. (2006). Child ADHD and personality/temperament traits of reactive and effortful control, resiliency, and emotionality. *Journal of Child Psychology and Psychiatry*, *47*(11), 1175–1183. https://doi.org/10.1111/j.1469-7610.2006.01629.x

Matthews, J. S., Marulis, L. M., & Williford, A. P. (2014). Gender processes in school functioning and the mediating role of cognitive self-regulation. *Journal of Applied Developmental Psychology*, *35*(3), 128–137. https://doi.org/10.1016/j.appdev.2014.02.003

Matthews, J. S., Ponitz, C. C., & Morrison, F. J. (2009). Early gender differences in self-regulation and academic achievement. *Journal of Educational Psychology*, *101*(3), 689–704. https://doi.org/10.1037/a0014240

Mazzocco, M. M. M., & Kover, S. T. (2007). A longitudinal assessment of executive function skills and their association with math performance. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *13*(1), 18–45. https://doi.org/10.1080/09297040600611346

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods*, *14*(2), 126–149. https://doi.org/10.1037/a0015857

McClelland, M. M., Cameron, C. E., Connor, C. M., 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., Geldhof, G. J., Cameron, C. E., & Wanless, S. B. (2015). Development and self-regulation. In R. M. Lerner (Ed.), *Handbook of child psychology and developmental science* (pp. 1–43). https://doi.org/10.1002/9781118963418.childpsy114

McClelland, M. M., Ponitz, C. C., Messersmith, E. E., & Tominey, S. (2010). Self-regulation. In R. M. Lerner (Ed.), *The handbook of life-span development*. https://doi.org/10.1002/9780470880166.hlsd001015

McCoy, D. C. (2019). Measuring young children's executive function and self-regulation in classrooms and other real-world settings. *Clinical Child and Family Psychology Review*, *22*(1), 63–74. https://doi.org/10.1007/s10567-019-00285-1

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review*, *106*(1), 3–19. https://doi.org/10.1037/0033-295X.106.1.3

Mischel, W., & Ayduk, O. (2004). Willpower in a cognitive-affective processing system: The dynamics of delay of gratification. In R. F. Baumeister & K. D. Vohs (Eds.), *Handbook of self-regulation: Research, theory, and applications* (pp. 99–129). The Guilford Press.

Mischel, W., Ayduk, O., Berman, M. G., Casey, B. J., Gotlib, I. H., Jonides, J., Kross, E., Teslovich, T., Wilson, N. L., Zayas, V., & Shoda, Y. (2011). 'Willpower' over the life span: Decomposing self-regulation. *Social Cognitive and Affective Neuroscience*, *6*(2), 252–256. https://doi.org/10.1093/scan/nsq081

Moeller, J. (2015). A word on standardization in longitudinal studies: Don't. *Frontiers in Psychology*, *6*, 1389. https://doi.org/10.3389/fpsyg.2015.01389

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., Houts, R., Poulton, R., Roberts, B. W., Ross, S., Sears, M. R., Thomson, W. M., & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, *108*(7), 2693–2698. https://doi.org/10.1073/pnas.1010076108

Montroy, J. J., Bowles, R. P., Skibbe, L. E., McClelland, M. M., & Morrison, F. J. (2016). The development of self-regulation across early childhood. *Developmental Psychology*, *52*(11), 1744–1762. https://doi.org/10.1037/dev0000159

Murray, K. T., & Kochanska, G. (2002). Effortful control: Factor structure and relation to externalizing and internalizing behaviors. *Journal of Abnormal Child Psychology*, *30*(5), 503–514. https://doi.org/10.1023/A:1019821031523

Natesan, P., Nandakumar, R., Minka, T., & Rubright, J. D. (2016). Bayesian prior choice in IRT estimation using MCMC and variational bayes. *Frontiers in Psychology*, *7*, 1422. https://doi.org/10.3389/fpsyg.2016.01422

Nigg, J. T. (2017). Annual Research Review: On the relations among self-regulation, self-control, executive functioning, effortful control, cognitive control, impulsivity, risk-taking, and inhibition for developmental psychopathology. *Journal of Child Psychology and Psychiatry*, *58*(4), 361–383. https://doi.org/10.1111/jcpp.12675

Oldehinkel, A. J. (2016). Editorial: Bayesian benefits for child psychology and psychiatry researchers. *Journal of Child Psychology and Psychiatry*, *57*(9), 985–987. https://doi.org/10.1111/jcpp.12619

Olson, S. L., Sameroff, A. J., Kerr, D. C. R., Lopez, N. L., & Wellman, H. M. (2005). Developmental foundations of externalizing problems in young children: The role of effortful control. *Development and Psychopathology*, *17*(1), 25–45. https://doi.org/10.1017/s0954579405050029

Ospina, R., & Ferrari, S. L. P. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, *56*(6), 1609–1623. https://doi.org/10.1016/j.csda.2011.10.005

Padilla-Walker, L. M., & Christensen, K. J. (2011). Empathy and self-regulation as mediators between parenting and adolescents' prosocial behavior toward strangers, friends, and family. *Journal of Research on Adolescence*, *21*(3), 545–551. https://doi.org/10.1111/j.1532-7795.2010.00695.x

Petersen, I. T., Bates, J. E., McQuillan, M. E., Hoyniak, C. P., Staples, A. D., Rudasill, K. M., Molfese, D. L., & Molfese, V. J. (2021). Heterotypic continuity of inhibitory control in early childhood: Evidence from four widely used measures. *Developmental Psychology*, *57*(11), 1755–1771. https://doi.org/10.1037/dev0001025

Petersen, I. T., Choe, D. E., & LeBeau, B. (2020). Studying a moving target in development: The challenge and opportunity of heterotypic continuity. *Developmental Review*, *58*, 100935. https://doi.org/10.1016/j.dr.2020.100935

Petersen, I. T., Hoyniak, C. P., McQuillan, M. E., Bates, J. E., & Staples, A. D. (2016). Measuring the development of inhibitory control: The challenge of heterotypic continuity. *Developmental Review*, *40*, 25–71. https://doi.org/10.1016/j.dr.2016.02.001

Petersen, I. T., & LeBeau, B. (2021). Language ability in the development of externalizing behavior problems in childhood. *Journal of Educational Psychology*, *113*(1), 68–85. https://doi.org/10.1037/edu0000461

Petersen, I. T., LeBeau, B., & Choe, D. E. (2021). Creating a developmental scale to account for heterotypic continuity in development: A simulation study. *Child Development*, *92*(1), e1–e19. https://doi.org/10.1111/cdev.13433

Petersen, I. T., Lindhiem, O., LeBeau, B., Bates, J. E., Pettit, G. S., Lansford, J. E., & Dodge, K. A. (2018). Development of internalizing problems from adolescence to emerging adulthood: Accounting for heterotypic continuity with vertical scaling. *Developmental Psychology*, *54*(3), 586–599. https://doi.org/10.1037/dev0000449

Petitclerc, A., Briggs-Gowan, M. J., Estabrook, R., Burns, J. L., Anderson, E. L., McCarthy, K. J., & Wakschlag, L. S. (2015). Contextual variation in young children's observed disruptive behavior on the DB-DOS: Implications for early identification. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *56*(9), 1008–1016. https://doi.org/10.1111/jcpp.12430

Posner, M. I., & Rothbart, M. K. (2000). Developing mechanisms of self-regulation. *Development and Psychopathology*, *12*(3), 427–441. https://doi.org/10.1017/S0954579400003096

Putnam, S. P., Rothbart, M. K., & Gartstein, M. A. (2008). Homotypic and heterotypic continuity of fine-grained temperament during infancy, toddlerhood, and early childhood. *Infant and Child Development*, *17*(4), 387–405. https://doi.org/10.1002/icd.582

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reed, R. G., Combs, H. L., & Segerstrom, S. C. (2020). The structure of self-regulation and its psychological and physical health correlates in older adults. *Collabra: Psychology*, *6*(1), 23. https://doi.org/10.1525/collabra.297

Revelle, W. (2020). *psych: Procedures for psychological, psychometric, and personality research*. Northwestern University, Evanston, Illinois. R package version 2.0.12. https://CRAN.R-project.org/package=psych

Roebers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review*, *45*, 31–51. https://doi.org/10.1016/j.dr.2017.04.001

Rothbart, M. K., & Bates, J. E. (2006). Temperament. In N. Eisenberg, W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology: Social, emotional, and personality development* (pp. 99–166). John Wiley & Sons, Inc.

Sasser, T. R., Bierman, K. L., & Heinrichs, B. (2015). Executive functioning and school adjustment: The mediational role of pre-kindergarten learning-related behaviors. *Early Childhood Research Quarterly*, *30*(Pt A), 70–79. https://doi.org/10.1016/j.ecresq.2014.09.001

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). *Woodcock-Johnson IV Tests of Cognitive Abilities*. Rolling Meadows, IL: Riverside.

Schrank, F. A., Wendling, B. J., Flanagan, D. P., & McDonough, E. M. (2018). The Woodcock–Johnson IV Tests of Early Cognitive and Academic Development. In *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 283–301). The Guilford Press.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge University Press. https://doi.org/10.1017/CBO9780511526817

Simpson, A., & Carroll, D. J. (2019). Understanding early inhibitory development: Distinguishing two ways that children use inhibitory control. *Child Development*, *90*(5), 1459–1473. https://doi.org/10.1111/cdev.13283

Spinrad, T. L., Eisenberg, N., Cumberland, A., Fabes, R. A., Valiente, C., Shepard, S. A., Reiser, M., Losoya, S. H., & Guthrie, I. K. (2006). Relation of emotion-related regulation to children's social competence: A longitudinal study. *Emotion*, *6*(3), 498–510. https://doi.org/10.1037/1528-3542.6.3.498

Stan Development Team. (2020b). *RStan: The R interface to Stan. R package version 2.21.0*. http://mc-stan.org/

Stan Development Team. (2020a). *RStan: The R interface to Stan. R package version 2.21.3*. http://mc-stan.org/

Stiles, J., & Jernigan, T. L. (2010). The basics of brain development. *Neuropsychology Review*, *20*(4), 327–348. https://doi.org/10.1007/s11065-010-9148-4

Sulik, M. J., Huerta, S., Zerr, A. A., Eisenberg, N., Spinrad, T. L., Valiente, C., Di Giunta, L., Pina, A. A., Eggum, N. D., Sallquist, J., Edwards, A., Kupfer, A., Lonigan, C. J., Phillips, B. M., Wilson, S. B., Clancy-Menchetti, J., Landry, S. H., Swank, P. R., Assel, M. A., & Taylor, H. B. (2010). The factor structure of effortful control and measurement invariance across ethnicity and sex in a high-risk sample. *Journal of Psychopathology and Behavioral Assessment*, *32*(1), 8–22. https://doi.org/10.1007/s10862-009-9164-y

Tiego, J., Bellgrove, M. A., Whittle, S., Pantelis, C., & Testa, R. (2020). Common mechanisms of executive attention underlie executive function and effortful control in children. *Developmental Science*, *23*(3), e12918. https://doi.org/10.1111/desc.12918

Twisk, J., de Boer, M., de Vente, W., & Heymans, M. (2013). Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. *Journal of Clinical Epidemiology*, *66*(9), 1022–1028. https://doi.org/10.1016/j.jclinepi.2013.03.017

Ursache, A., Blair, C., & Raver, C. C. (2012). The promotion of self-regulation as a means of enhancing school readiness and early achievement in children at risk for school failure. *Child Development Perspectives*, *6*(2), 122–128. https://doi.org/10.1111/j.1750-8606.2011.00209.x

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2014). A gentle introduction to bayesian analysis: Applications to developmental research. *Child Development*, *85*(3), 842–860. https://doi.org/10.1111/cdev.12169

Whedon, M., Perry, N. B., Curtis, E. B., & Bell, M. A. (2021). Private speech and the development of self-regulation: The importance of temperamental anger. *Early Childhood Research Quarterly*, *56*, 213–224. https://doi.org/10.1016/j.ecresq.2021.03.013

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838. https://doi.org/10.2307/1912934

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10–18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

Wiebe, S. A., Espy, K., & Charak, D. (2008). Using confirmatory factor analysis to understand executive control in preschool children: I. latent structure. *Developmental Psychology*, *44*, 575–587. https://doi.org/10.1037/0012-1649.44.2.575

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-old children. *Journal of Experimental Child Psychology*, *108*(3), 436–452. https://doi.org/10.1016/j.jecp.2010.08.008

Willoughby, M. T., Kuhn, L. J., Blair, C. B., Samek, A., & List, J. A. (2017). The test–retest reliability of the latent construct of executive function depends on whether tasks are represented as formative or reflective indicators. *Child Neuropsychology*, *23*(7), 822–837. https://doi.org/10.1080/09297049.2016.1205009

Willoughby, M. T., Kupersmidt, J., Voegler-Lee, M., & Bryant, D. (2011). Contributions of hot and cool self-regulation to preschool disruptive behavior and academic achievement. *Developmental Neuropsychology*, *36*(2), 162–180. https://doi.org/10.1080/87565641.2010.549980

Zelazo, P. D., & Cunningham, W. A. (2007). Executive function: Mechanisms underlying emotion regulation. In J. J. Gross (Ed.), *Handbook of emotion regulation* (pp. 135–158). The Guilford Press.

Zhou, Q., Chen, S. H., & Main, A. (2012). Commonalities and differences in the research on children's effortful control and executive function: A call for an integrated model of self-regulation. *Child Development Perspectives*, *6*(2), 112–121. https://doi.org/10.1111/j.1750-8606.2011.00176.x

Zimmermann, P., & Iwanski, A. (2014). Emotion regulation from early adolescence to emerging adulthood and middle adulthood: Age differences, gender differences, and emotion-specific developmental variations. *International Journal of Behavioral Development*, *38*(2), 182–194. https://doi.org/10.1177/0165025413515405

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hosch, A., Oleson, J. J., Harris, J. L., Goeltz, M. T., Neumann, T., LeBeau, B., Hazeltine, E., & Petersen, I. T. (2022). Studying children's growth in self-regulation using changing measures to account for heterotypic continuity: A Bayesian approach to developmental scaling. *Developmental Science*, *25*, e13280. https://doi.org/10.1111/desc.13280

Supplementary Appendix S1. Description of Participants.

There were no significant differences in the child's age (*t*[71.75] = -1.71, *p* = .091) or sex (χ^2[1] = 0.12, *p* = .728) between the participating families versus the screened but non-participating families. There were also no significant differences in the child's ethnicity or the recruitment location, though some frequencies in the cross-tabulation cells were sparse to evaluate this. Because we suspended data collection for 14 months during the COVID-19 pandemic, several prospective participants aged out of their eligibility window (i.e., 36, 45, 54, or 63 months of age) and were not able to be enrolled. In addition, many prospective participants have not yet aged into their eligibility window for subsequent timepoints or are currently negotiating scheduling. In addition to the lack of statistically significant differences between participating families versus the screened but non-participating families, effect sizes of differences were small. Thus, it does not appear that the lack of apparent differences was due to insufficient power. In sum, we do not have strong evidence to suggest that the participating and non-participating families differed in substantial ways.

Among participating primary caregivers (*n* = 109) and parenting partners (*n* = 103), 96.5% were biological parents, 1.4% were adoptive parents, 1.0% were stepparents, and 1.0% were grandparents. For simplicity, we refer to the primary caregiver and parenting partner as parents. Of secondary caregivers (*n* = 141), 44.0% were teachers, 24.0% were daycare providers, 27.4% were relatives, and 2.9% were babysitters. Participants included more primary caregivers than children because the identified primary caregiver changed across time for some children. The composition of marital status among primary caregivers was married (84.6%), remarried (1.8%), separated (1.3%), divorced (3.5%), and single or never married (8.8%). The composition of parents' educational attainment included: completed some high school (1.7%), completed high school (5.7%), completed some college (14.4%), Associate’s degree (9.2%), Bachelor’s degree (31.4%), Master’s degree (23.1%), professional school degree (8.3%), and doctoral degree (6.1%). The composition of socioeconomic status among participants is described in the section on Covariates in Supplementary Appendix S3. Compared to the U.S. population, participants in the sample were somewhat more likely to be Non-Hispanic White, married, be middle or upper class, and have a college or graduate degree. Participant demographics were broadly reflective of the surrounding area.

Supplementary Appendix S2. Description of missing data.

The extent of missingness for model variables is provided in Supplementary Table S1. Among possible participant-by-wave instances, 31.7% had missing scores because the child was not yet eligible for a given wave. Among eligible participant-by-wave instances, approximately 79% had self-regulation scores (see Supplementary Table S1). Among missing lab visits at a given wave for which the child reached eligibility, reasons for missingness included: not interested (12.1%), too busy (12.9%), moved/relocated (0.9%), unable to contact (14.7%), COVID-19 (50.0%), and other (9.4%). Thus, over half of missing instances were due to the coronavirus (COVID-19) pandemic or to not yet being eligible. We suspended lab visits for 14 months during the COVID-19 pandemic (March 2020 – April 2021). Thus, we have limited scores on performance-based assessments at participants' fourth timepoints because many occurred during the COVID-19 pandemic shutdown. However, we were able to collect online questionnaires from some families during the pandemic.

Tests of systematic missingness revealed no significant differences in missingness as a function of the child's ethnicity or externalizing problems. However, data were more likely to be missing for boys (compared to girls; $\chi^2[1] = 4.33$, $p = .037$) and were marginally more likely to be missing for children from lower socioeconomic status (SES) families ($t[80.40] = 1.81$, $p = .074$). Effect sizes of differences were small. In addition, older children were more likely to be missing self-regulation scores than younger children ($t[17.87] = -12.75$, $p < .001$), likely due to some attrition.

Supplementary Appendix S3. Description of Measures.

Measures

Self-regulation. Measures of self-regulation included 15 laboratory tasks and two questionnaires. Except for computer-scored tasks (Fish/Sharks and stop-signal tasks) and Token/Bead Sort, which was scored by counting tokens/beads in containers, children's performance on tasks was scored after the lab visit from video recording. All scored cases were double coded to evaluate inter-rater reliability via intraclass correlation. Raters met to potentially resolve any large discrepancies between raters' codes. Scores were averaged across raters. For development scaling, scores of each self-regulation measure were converted to proportion of maximum (POM) scores to have the same possible range (0–1), with higher scores reflecting greater self-regulation. For measures that had a minimum and maximum possible score, the POM score reflected the proportion of the maximum possible score. For measures that did not have a minimum or maximum possible score (stop-signal task and Token/Bead Sort), the POM score reflected the proportion of the maximum *observed* score. POM scores were calculated as: $\frac{\text{score} - \text{minimum}}{\text{maximum} - \text{minimum}}$, where minimum and maximum were the minimum and maximum possible score or the minimum and maximum observed score. Tasks (Token/Bead Sort; stop-signal task) and questionnaires (Behavior Rating Inventory of Executive Function) were adapted to accommodate the developmental capacity of the child and the changing expression of self-regulation with age.

***Bear/Dragon*.** Bear/Dragon (Kochanska et al., 1996) is a go/no-go task that assesses children's inhibitory control and set shifting. It involves activation on a subset of go trials and inhibition on a subset of no-go trials, based on the cue (i.e., puppet), with a rule reversal. The child was asked to follow instructions from a bear puppet, and to ignore instructions from a dragon puppet. The child completed three go and three no-go practice trials and was reminded of the rule if they failed a trial. Next, they were presented with 12 mixed test trials, including six go (i.e., bear) trials and six no-go (i.e., dragon) trials. Subsequently, the experimenter changed the rules and instructed the child to now follow the dragon's directions but ignore the bear. After six practice trials, the child completed a second set of 12 test trials (6 go trials and 6 no-go trials) in a pseudo-random order. Each no-go trial was scored from 1 to 4 (1 = full commanded movement, 2 = partial movement, 3 = wrong movement, and 4 = no movement). Scoring was reversed for go trials, consistent with Carlson and Moses (2001). Scores were averaged across trials within condition (no-go versus go). Because children could receive a high score on no-go trials by performing no action, we examined the degree to which children inhibited a response on no-go trials and activated a response on go trials. Consistent with Eisenberg et al. (2013), a composite of children's inhibition was computed by multiplying mean scores from inhibition (no-go) and activation (go) trials (1–16). Therefore, children who activated a behavior on go trials but inhibited on no-go trials received the highest scores, whereas children who never activated (or always activated) a behavior received low scores. Final scores were converted to a proportion of the maximum possible score. Higher scores reflected greater inhibitory control.

Day/Night. Day/Night (Gerstadt et al., 1994) assesses inhibitory control by inhibition of a prepotent association (that the sun is associated with daytime and the moon is associated with nighttime) and generation of a competing response. In this task, the child was shown two kinds of cards: one with a picture of a sun on a white background and the other with a picture of a moon on a black background. The child was instructed to say "day" when they see the card with the black moon and say "night" when they see the card with the yellow sun. The task began with two practice trials, during which the experimenter praised the child for correct responses. If the child responded incorrectly to either practice trial, the experimenter reminded them of the rules and repeated the trials. After completing the practice trials, the child was presented with 16 test trials, eight of each word, in a fixed, quasi-random order. During the 16 test trials, the experimenter did not provide feedback. Each trial was scored incorrect (0), initially incorrect, but changed to correct (1), or correct (2), consistent with Kochanska et al.'s (1997) scoring of other inhibitory tasks. The final score was the average score across trials (0–2). Final scores were converted to a proportion of the maximum possible score. Higher scores reflected greater inhibitory control.

***Fish/Sharks*.** Fish/Sharks (Wiebe et al., 2012) is a go/no-go task that assesses inhibitory control. The task was administered on a computer using E-Prime software (version 2.0.10.356; Schneider et al., 2012). During the task, the child was shown cartoon images of fish (go stimuli) and sharks (no-go stimuli) on a touch screen. Stimuli included ten fish and three sharks. However, on any given trial, only one fish or one shark was presented. The child was instructed to touch the fish to catch the fish in their net and not to touch the sharks because the sharks are too big and would break their net. On go trials in which the child touched the fish, positive feedback was presented: an image of the fish in the net and pleasant bubble sounds. On no-go trials in which the child touched the shark, an image of the shark breaking the net and an unpleasant buzzer sound was presented. No feedback was given if the child successfully inhibited (except during practice trials).

The task began with four practice blocks, each with eight practice trials, in the following order: go trials only, no-go trials only, go-trials only, and mixed (i.e., both go and no-go trials) practice block. The experimenter gave feedback during the practice trials. After the child successfully completed the practice trials, the test trials began. The test trials consisted of 80 trials: 60 go trials and 20 no-go trials. The task was split into ten blocks of eight trials. Each block included six go trials and two no-go trials that were randomly presented. The stimuli (i.e., fish or sharks) were presented for a maximum of 3000 milliseconds or until the child touched the screen. Feedback stimuli were presented after the child touched the screen and were displayed for 750 ms. The overall inter-stimulus interval was 1500 ms. The experimenter provided rule reminders during the test trials but did not provide corrective feedback.

Behavioral responses that occurred less than 200 ms after stimulus onset were discarded from analyses because this would be too rapid for the child to have responded deliberately to the target stimulus. A composite of children's inhibition was computed by multiplying the proportion of correct inhibition (no-go) trials by the proportion of correct activation (go) trials, consistent with Eisenberg et al. (2013). Children who activated a behavior on go trials but inhibited on no-go trials received the highest scores, whereas children who never activated (or always activated) a behavior received low scores. Final scores were converted to a proportion of the maximum possible score. Higher scores reflected greater inhibitory control.

Gift Delay. Gift Delay (Kochanska et al., 2000) is a delay-of-gratification task that assesses children's motivational self-regulation. The child was presented with a gift and asked to refrain from touching and looking at it until the experimenter gave permission. The experimenter praised the child for their cooperation in the previous tasks and promised a surprise gift. However, before receiving the gift, the child was instructed to turn away from the table, so the experimenter could wrap the gift out of their sight. While wrapping the gift ("wrapping" waiting period), the experimenter created loud noises by rustling the tissue paper and shaking the gift bag. After one minute, the experimenter allowed the child to turn around and look at the gift bag. The experimenter announced that they would leave the room to find a bow for the gift. The child was instructed not to look inside the gift bag until the experimenter returned ("gift-in-bag" waiting period). After three minutes, the experimenter re-entered the room and prompted the child to open the gift.

The task was coded along a six-tier hierarchy of how the child interacted with the gift during the one-minute "wrapping" waiting period and the three-minute "gift-in-bag" waiting period. The hierarchy ranged from the most obedient behavior to the most disobedient behavior (1 = never looks at and never touches gift; 2 = looks at gift bag; 3 = touches gift bag; 4 = looks inside gift bag; 5 = touches paper or toy inside gift bag; 6 = opens gift), adapted from Kochanska et al. (2000). Because the experimenter was in the room during the wrapping waiting period but not during the gift-in-bag waiting period, the frequencies of the various disobedient behaviors differed across the two waiting periods. Therefore, to adequately capture variability of responses, we assigned children's scores to different values for each condition, including a wrap score (1 = looks inside gift bag, touches paper or toy inside git bag, or opens gift; 2 = touches gift bag; 3 = looks at gift bag but does not stay in their seat the entire time; 4 = looks at gift bag, and stays in their seat the entire time; 5 = never looks at and never touches gift or gift bag) and a gift score (1 = opens gift; 2 = touches paper or toy inside gift bag; 3 = looks inside gift bag; 4 = touches gift bag; 5 = never looks inside gift bag and never touches gift or gift bag but does not stay in their seat the entire time; 6 = never looks inside gift bag and never touches gift or gift bag, and stays in their seat the entire time). Wrap scores were correlated with gift scores (*r* = .19, *p* = .011)**.** The child's wrap score was averaged with their gift score. Final scores were converted to a proportion of the maximum possible score. Higher scores reflected greater delay of gratification.

Grass/Snow. Grass/Snow (Carlson & Moses, 2001) assesses inhibitory control by inhibition of a prepotent association (that the word "grass" is associated with the color green and the word “snow” is associated with the color white) and generation of a competing response. In the task, the child was instructed to touch a white square when they hear the word “grass” and a green square when they hear the word “snow.” The task began with several practice trials, during which the experimenter praised the child for correct responses. If the child responded incorrectly to a practice trial, the experimenter reminded the child of the rules and repeated the trials. Following these practice trials, the child was presented with 12 trials, six of each word, in a fixed, quasi-random order, and each trial was scored either correct (1) or incorrect (0), consistent with Carlson and Moses (2001). Final scores were averaged across trials (0–1), which reflected a proportion of maximum possible score. Higher scores reflected greater inhibitory control.

Hand Game. Hand game (Luria et al., 1964) assesses inhibitory control. In this task, the child was instructed to either point a finger or make a fist, in response to the experimenter’s hand movement. During the six initial imitation checks, the child copied the experimenter’s hand movements to ensure the child had the motor abilities to complete the task. Subsequently, the child was asked to point a finger when the experimenter made a first, and to make a fist when the experimenter pointed a finger. The task began with two comprehension check trials, one for each movement, followed by six practice trials. The experimenter praised the child for correct trials. If the child responded incorrectly, the experimenter reminded the child of the rules and repeated the trial. After completing the practice trials, the child was presented with 15 test trials, in a fixed, quasi-random order. During these test trials, the experimenter did not provide feedback. Each trial was scored incorrect (0), initially incorrect, but changed to correct (1), or correct (2), consistent with Kochanska et al.'s (1997) scoring of other inhibitory tasks. Scores were averaged across all trials (0–2). Final scores were converted to a proportion of the maximum possible score. Higher scores reflected greater inhibitory control.

Knock/Tap. Knock/Tap (Klenberg et al., 2001) assesses inhibitory control and shifting and consists of two parts. Prior to starting the task, two imitation trials were administered to ensure the child had the motor abilities to complete the task. During these imitation trials, the child copied how the experimenter knocked or tapped the table. The child was then instructed to knock the table, whenever the experimenter tapped, and to tap the table whenever the experimenter knocked. After two comprehension checks and two practice trials, 15 pseudo-random test trials were administered. In the second part of the task, the instructions changed. The child was instructed to make a side fist when the experimenter knocked, and to knock when the experimenter made a side fist. However, when the experimenter tapped the table, the child was instructed to do nothing. After six practice trials, 15 test trials were administered. During test trials, the experimenter did not provide feedback. Each trial was scored incorrect (0), initially incorrect, but changed to correct (1), or correct (2), consistent with Kochanska et al.'s (1997) scoring of other inhibitory tasks. Scores were averaged across trials (0–2). Final scores were converted to a proportion of the maximum possible score. Higher scores reflected greater inhibitory control.

Less is More. Less is More is a motivationally salient symbolic representation task that assesses affective (“hot”) inhibitory control (Carlson et al., 2005). The child chose a preferred treat from two options, white marshmallows and uniformly colored jelly beans. The preferred treats were pre-bagged in transparent bags with some bags containing two treats and others containing five treats. The child was asked if they prefer the bag of two treats or five treats. Children who preferred the two treat bags at the beginning of the trial were excluded. In front of the child were two bowls, one of which had a “naughty monkey” puppet, and the other bowl was the child’s bowl. The child was told that “the monkey wants all the treats for himself.” On each trial, two bags are presented to the child: one bag with five treats and one bag with two treats. The child was instructed to point to a bag among the two bag options presented. The child was instructed that the bag they point to goes to the monkey’s bowl, and that they receive the treats in the other bag (i.e., the bag they did not point to). Each time the child chose a bag, the experimenter put the bag the child chose in the monkey’s bowl, and the other bag in the child’s bowl. After up to three comprehension check trials with corrective feedback, there were eight test trials in the first trial set. The monkey was then moved to the opposite bowl to avoid a side bias. Then, another comprehension check and eight more trials were administered with the same rules as the first trial set. Responses were scored as: 0 = child points to large treats bag; 1 = child initially points to the large treats bag, then changes to the small treats bag; 2 = child points to the small treats bag, consistent with Kochanska et al.'s (1997) scoring of other inhibitory tasks. Scores were averaged across 16 test trials (0–2). Final scores were converted to a proportion of the maximum possible score. Higher scores reflected greater affective inhibitory control.

Peg Tapping. Peg Tapping (Luria et al., 1964) assesses inhibitory control. The child observed sequences of a specific number of pencil taps on a table (either one or two) and was instructed to tap a pencil the opposite number of times of what they observed. The experimenter explained the rules: when the experimenter taps the pencil once and then hands the pencil to the child, the child is to tap the pencil twice. When the experimenter taps the pencil twice, the child is to tap the pencil once. The child received two practice trials and then received 16 test trials in which the experimenter followed a fixed, quasi-random order to tap once or twice. The child was given corrective feedback on the practice trials but not the test trials. Trials were scored correct (1) or incorrect (0). Final scores were averaged across trials, which reflected a proportion of maximum possible score. Higher scores reflected greater inhibitory control.

Self-imposed waiting task. The self-imposed waiting task (Metcalfe & Mischel, 1999) is a delay-of-gratification task that assesses children's motivational self-regulation. The task is designed to assess children's ability or preference to resist the temptation of immediate gratification in favor of a more motivationally salient reward at a later time. In the task, the child was presented with a bell, and the experimenter explained that ringing the bell would bring the experimenter back into the room at any time. The child completed several practice rounds, in which the experimenter left the room and re-entered when the child rang the bell. Following these practice rounds, multiple options of treats were introduced: chocolate chips, bear-shaped graham crackers, and oyster crackers. The child chose a most preferred and a least preferred treat, as suggested by Neuenschwander and Blair (2017). Two plates of treats were presented: a plate with a large portion of their most preferred treat, and a plate with a small portion of their least preferred treat, as adapted from Mischel et al. (1972), Razza and Raymond (2013), and Duckworth et al. (2013). We placed both plates of treats to be present during the task because research has shown that waiting times are shorter, on average, when both the most preferred and least preferred treats are present, compared to when only one or neither is present (Peake, 2017).

The experimenter then presented the child with the instructions for the waiting game. The child was informed that the experimenter would leave the room for a period, and the child would be allowed to eat the big plate of preferred treats if they waited for the entire period until the experimenter returned. The child was told that they could, at any time, ring the bell to bring the experimenter back into the room. If they did so, however, they could not eat the big plate of the preferred snack; if they chose to ring the bell, they could eat only the small plate of the least preferred snack. The full duration of the waiting period was recorded for each child, beginning when the experimenter left the room and ending either when the bell was rung, the treat was prematurely consumed, or the seven-minute period had elapsed. The child's score was the duration of their waiting period in seconds. Final scores were converted to a proportion of the maximum possible score. Higher scores reflected greater delay of gratification.

Shape Stroop. Shape Stroop (Kochanska et al., 2000) assesses children's perceptual inhibitory control. The task assessed the child's ability to identify a picture of a small fruit embedded within a picture of a different, larger fruit. To verify that the child knew the names of the fruits in the pictures, the child was first presented three pictures, each containing one large fruit: an apple, banana, or orange. In the first three trials, the child was asked to point to a large fruit (e.g., the large apple). After successfully identifying these three fruits, the child was presented with three new pictures, each containing a small fruit embedded within a different, larger fruit image (e.g., a small banana embedded within a larger apple image). The following three trials, the child was instructed to point to a small fruit (e.g., the small banana). Trials were scored from 0 to 2 (0 = incorrect, 1 = initially incorrect, but changed response to correct, 2 = correct; Kochanska et al., 2000). Scores were averaged across the three small fruit trials (0–2). Final scores were converted to a proportion of the maximum possible score. Higher scores reflected greater perceptual inhibitory control.

Simon Says. Simon Says (Strommen, 1973) assesses children's inhibitory control in response to verbal and motor cues. The task involved a series of activation (i.e., "go") and inhibition (i.e., "no-go") trials, in which the child was instructed to inhibit their behavioral response to instructions unless the instructions are accompanied by a verbal cue. The child was presented with a series of instructions to perform simple motor actions (e.g., clap your hands, stomp your feet) and was told to perform the action only if the instructions are preceded by the phrase “Simon Says.” The child completed two go practice trials and two no-go practice trials, followed by 20 test trials, including ten go trials and ten no-go trials, presented in a fixed, pseudo-random order. Each no-go trial was scored from 1 to 4 (1 = full commanded movement, 2 = partial movement, 3 = wrong movement, and 4 = no movement), consistent with Carlson and Moses (2001) scoring of a simplified version of Simon Says (Bear/Dragon); scoring was reversed for go trials. Scores were averaged across trials within condition (no-go versus go; ranged 1–4). Because children could receive a high score on no-go trials by simply not responding, a composite score of children’s inhibitory control was computed by multiplying mean scores from ten go trials and ten no-go trials, consistent with Eisenberg et al. (2013). Children who inhibited behavior across all trials thus received a lower score compared to children who correctly inhibited behavior across inhibition (no-go) trials and activated behavior across activation (go) trials. Final scores were converted to a proportion of the maximum possible score. Higher scores reflected greater inhibitory control.

***Snack Delay*.** Snack Delay (Kochanska et al., 1996) is a delay-of-gratification task that assesses children’s motivational self-regulation. The task assesses children’s ability to suppress a dominant behavioral response and perform a subdominant response when presented with a highly motivating stimulus. The child was presented with a chosen treat (e.g., M&Ms) placed under a clear plastic cup and a placemat displaying an image of handprints. The child was instructed to keep their hands on the placemat and to refrain from picking up the cup and eating the treat until the experimenter rings a bell. The child completed a practice trial, followed by five test trials of differing delay times (5 s, 10 s, 20 s, 30 s, and 15 s in length) in which the experimenter lifted the bell halfway through the trial but did not ring it until the full time elapsed. Thus, each trial contained two segments (i.e., bell on table and bell lifted). The child was reminded of the rules before each trial. Each segment was scored from 0 to 3, with higher scores indicating greater delay performance: bell-on-table segment (0 = eats snack before bell is lifted, 1 = touches snack before bell is lifted, 2 = touches glass and/or bell before bell is lifted; 3 = waits until bell is rung), and bell lifted segment (0 = eats snack before bell is rung, 1 = touches snack before bell is rung, 2 = touches glass and/or bell before bell is rung; 3 = waits until bell is rung). Children who did not eat the treat at any point throughout the task were excluded from analyses. An additional point was awarded if the child's hands remained on the placemat throughout the segment, resulting in a final segment score of 0 to 4, as adapted from Kochanska et al. (2000). Scores were averaged across all ten segments. Final scores were converted to a proportion of the maximum possible score. Higher scores reflected greater delay of gratification.

Stop-signal task. The stop-signal task is a widely used experimental procedure to assess the ability to inhibit inappropriate actions (Verbruggen et al., 2019). The Food Finder stop-signal task was adapted from Berger et al. (2013) to be more appropriate for children as young as three years of age with child-friendly stimuli, an engaging storyline, animations, touchscreen, and a progress bar. Children performed a two-alternative forced choice task, but on some trials, they were given a cue (stop signal) to withhold responding. If the stop signal appeared too late after the go stimulus, children were unable to withhold the response. Latency of the stop signal after go stimulus onset (stop-signal delay [SSD]) was manipulated to determine a child's speed of response inhibition.

The task included three blocks that followed the same structure: presentation of go stimuli, practice go trials, presentation of stop signal, mixed practice trials, and test trials. Each trial began with a flickering star in the center of the screen that served as a fixation point. In Block 1, trials included a picture of a green food (e.g., lime) or purple food (e.g., grapes) in the middle of the screen. On the bottom of the screen was a picture of a green goat and purple pig. The child was told to give purple food to the purple pig and green food to the green goat by touching the animal on the screen. The child was told to touch the purple pig when they see purple food and to touch the green goat when they see green food. The child then completed the practice go trials and experimenters provided praise for correct responses. After completing the practice go trials, the child was shown a cartoon wizard and was told that wizard will try to trick them and turn the food into a car. On stop trials, the food and animals were shown, and after some delay (i.e., SSD) the food and animals were replaced by a car. The child was instructed not to feed cars to the animals and not to touch the screen when they saw a car. The child was instructed to go as fast as they can. The child then completed mixed practice trials, i.e., both go and stop trials. After the mixed practice trials, the child completed the test trials which consisted of 60 trials in each block: 42 go trials and 18 stop trials.

The task used a staircase dynamic-tracking paradigm that adjusted the SSD based on the child's performance on previous stop trials. The algorithm adjusting the SSD attempted to obtain a 50% error rate on stop trials, which helped normalize task difficulty across ages. The SSD was set at 400 ms for the first trial of Block 1 so the task would be relatively easy in the beginning and become more challenging over time. The delay modification after each stop trial was 100 ms during Block 1 and was 50 ms in Blocks 2 and 3. The delay modification was higher in Block 1 than Blocks 2 and 3 to converge upon the 50% error rate more quickly. If the participant successfully inhibited on a stop trial, the delay modification was added to the SSD on the next stop trial to make stopping more difficult. If the participant failed to inhibit on a stop trial or if they responded before the stop signal, the delay modification was subtracted from the SSD on the next stop trial to make stopping easier. The running SSD at the end of each block carried forward to the next block.

The trial stimuli (i.e., food, animals, and cars) were presented for a maximum of 5000 ms or until the child touched the screen. Auditory feedback lasting ~540–700 ms was provided after every trial. Feedback was a “yippee” sound for all correct trials: correct responses on go trials and successful omissions on stop trials. For correct responses on go trials, animation showed the food moving toward the selected animal. Feedback was a “hmm” sound for all incorrect trials: omission errors on go trials, incorrect categorizations on go trials—i.e., touching the picture of the wrong animal, and commission errors on stop trials.

To reduce habituation, the animals and foods changed in Blocks 2 and 3. In Block 2, the child was told to give orange food to the orange owl and red food to the red rabbit. In Block 3, the child was told to feed blue food to the blue bird and pink food to the pink penguin. The cartoon wizard and cars were kept the same for both Blocks 2 and 3. There were three foods of each color. In Blocks 2 and 3, the children completed the test trials only, for a total of 180 trials (126 go and 54 stop trials). Again, feedback was provided on every trial. Stimuli were presented via E-Prime software.

We performed several processing steps to ensure data were high-quality. The length and difficulty of task blocks caused some children to fail to perform the task for some subsets of trials. We attempted to identify these subsets of children and trials to retain as many children and trials in the analyses as possible while eliminating trials that did not tap response inhibition processes and children who had insufficient valid trials. No algorithm will be perfectly accurate in adjudicating valid responding, but the following criteria were adopted to restrict the analysis to trials in which the child appeared to be performing the task as instructed while allowing for temporary lapses. The same criteria were applied to all children.

First, responses that occurred less than 200 ms after go stimulus onset were discarded from analyses because this would be too rapid for the child to have responded deliberately to the target stimulus. We excluded subsets of trials during which the child appeared to be temporarily deviating from the instructions but later returned to the task. In some cases, children consistently delayed their response to wait for the stop signal, causing the SSD to become so long that it was no longer relevant for task performance. These subsets of trials were identified by sequences of six or more stop trials in which the child responded before the stop signal appeared. For these trial subsets, we kept only those data prior to the first instance of responding before the stop signal (in that sequence of six consecutive stop trials), and we retained trials after the child had three consecutive stop trials in which they did not respond before the stop signal. If the child had a sequence of trials in which they appeared not to be participating (i.e., they did not respond on four or more consecutive go trials), we kept only those trials prior to their first missed go trial in that sequence of consecutive missed go trials. If the child started participating again, as operationalized by three failed stops in a sequence of six stop trials, we retained the subsequent trials.

We excluded children at a given measurement occasion who had insufficient valid trials due to excessive use of the strategy of delaying their response to wait for the stop signal. We set the threshold for insufficient trials due to this strategy as the child having 20% or more of their go trials in which their reaction times was shorter than the running SSD (i.e., the SSD at that point in the task). We also excluded children who did not have any failed stop trials, indicating that they requested infrequently or after a long delay. In addition, we excluded children who intentionally touched the stop signal (thus not following the rules), resulting in an unreasonably quick SSD. We set this threshold to exclude children whose mean SSD was less than 100 ms.

We operationalized response inhibition as the stop-signal reaction time (SSRT). The SSRT was calculated as the median reaction time on correct go trials minus the mean SSD from Blocks 2 and 3. Block 1 was not included in the calculation to allow the algorithm time to converge upon a 50% error rate on stop trials. Cases were excluded if the SSRT was negative (i.e., the median go reaction time was faster than the mean SSD). Final scores were converted to a proportion of the maximum observed score and were reverse scored. Higher scores reflected greater inhibitory control.

Token/Bead Sort. Token/Bead Sort (Goldsmith et al., 1999) is a behavioral performance task that assesses children's attentional self-regulation ability in a low-stimulation, academic-like task. The child was instructed to complete a sorting task of developmentally appropriate difficulty for an extended period of time. The child's sustained attention was assessed by their progress toward task completion (as an indicator of continued engagement) during the time period. The child was asked to sort a large pile of either multi-colored tokens (at ages 36–54 months) or small multi-colored beads (at ages 63–90 months) into separate containers based on color. The number of tokens and beads was preselected to ensure that the child would have difficulty finishing the sorting task in the allotted time. After the instructions were administered, the child was left alone in a room with the tokens/beads for three minutes. After the task was completed, the number of tokens in the correct and incorrect containers were totaled (tokens remaining unsorted were excluded). The child's score was computed by subtracting the number of incorrectly sorted tokens from the number of correctly sorted tokens. If the child sorted more tokens incorrectly than correctly (i.e., the final score was negative), which might indicate random token sorting, the score was set to zero. Final scores were converted to a proportion of the maximum observed score. Higher scores reflected greater sustained attention to the task and greater attentional regulation.

Questionnaires.

Behavior Rating Inventory of Executive Function. The Behavior Rating Inventory of Executive Function (BRIEF) assesses children's executive functioning within the context of their everyday environment. Two versions were used based on the child's age. Parents completed the BRIEF–Preschool Version (BRIEF–P; Gioia et al., 1996) if the child was 3–5 years old or the BRIEF–2 (Gioia et al., 2015) if the child was 6–7 years old. Scores on the Global Executive Composite were used for both questionnaires' versions. This score was composed of the Inhibit, Shift, Emotional Control, Working Memory, and Plan/Organize subscales for the BRIEF–P and the Inhibit, Self-Monitor, Shift, Emotional Control, Initiate, Working Memory, Plan/Organize, Task-Monitor, and Organization of Materials subscales for the BRIEF–2. The three additional subscales used in the BRIEF–2 (compared to the BRIEF–P) reflect developmental changes in executive functioning from ages 3–7 years, including higher-order regulatory processes. Items were also adapted to account for normative increases in working memory skills with age and changes in activities and contexts, such as chores and homework. For example, one item on the BRIEF–P is "When given *two* things to do, remembers only the first or last", whereas the corresponding item on the BRIEF–2 is "When given *three* things to do, remembers only the first or last." Sixty-three items were rated on a 3-point scale (1 = never, 2 = sometimes, 3 = often) in terms of how often, in the last six months, the child's behavior had been a problem. To account for missing responses in the sum score, scores were averaged across items and then multiplied by the number of items. Scores were converted to a proportion of the maximum (POM) possible score. Scores were then reverse scored so that higher scores reflected greater executive functioning. Mothers' and fathers' ratings on the Global Executive Composite were significantly correlated (*r*[95] = .35, *p* < .001). Age and sex norm-referenced *T*-scores had a mean of 51.42 (*SD* = 10.38).

Children's Behavior Questionnaire. The Children's Behavior Questionnaire (CBQ) assesses children's temperament (i.e., reactivity and regulation). Two versions were used based on the rater type. Parents completed the CBQ (Putnam & Rothbart, 2006). Secondary caregivers completed the CBQ–Teacher Short Form (CBQ–TSF, Teglasi et al., 2015). The CBQ and CBQ–TSF instruments consist of three general temperament dimensions: Negative Affectivity, Surgency/Extraversion, and Effortful Control. We used scores from the Effortful Control scale (CBQ: 47 items; CBQ–TSF: 26 items), which consists of the Attentional Focusing, Inhibitory Control, Low Intensity Pleasure, and Perceptual Sensitivity subscales. Items were rated on 7-point Likert scale (1 = extremely untrue, 2 = quite untrue, 3 = slightly untrue, 4 = neither true nor untrue, 5 = slightly true, 6 = quite true, 7 = extremely true). Scores were averaged across items. Scores were converted to a proportion of the maximum possible score. Higher scores reflected greater effortful control. Mothers' ratings on the Effortful Control scale were associated with ratings by fathers (*r*[104] = .51, *p* < .001) and secondary caregivers (*r*[103] = .38, *p* < .001). Fathers' ratings were associated with ratings by secondary caregivers (*r*[72] = .43, *p* < .001).

School Readiness

Woodcock Johnson IV – Tests of Achievement. The Woodcock Johnson IV – Tests of Achievement (Schrank et al., 2014, 2018) assess academic achievement. Children completed two subtests to assess their early (pre-)reading and math skills: Letter-Word Identification and Applied Problems, respectively. Letter-Word Identification (78 items) assesses word identification skills and reading-writing ability. The child was asked to identify letters and eventually asked to read aloud individual words. Applied Problems (56 items) assesses quantitative knowledge ability. The child was asked to analyze and solve applied math problems. Items were scored on accuracy (1 = correct, 0 = incorrect). Raw scores (i.e., the number of correct responses) we used. Higher scores reflected better school readiness. Age norm-referenced standard scores had a mean of 99.98 (*SD* = 14.37) and 103.12 (*SD* = 16.16) for reading and math skills, respectively.

Externalizing Behavior

Achenbach System of Empirically Based Assessment. The Achenbach System of Empirically Based Assessment (ASEBA) assesses children's emotional and behavioral problems. Items were rated on a 3-point Likert scale according to how well the item described the child (0 = not true, 1 = somewhat or sometimes true, 2 = very true). Multiple versions were used based on the child's age and rater type. Parents completed the Child Behavior Checklist 1.5–5 (CBCL 1.5–5; Achenbach & Rescorla, 2000) if the child was 3–5 years old or the Child Behavior Checklist 6–18 (CBCL 6–18; Achenbach & Rescorla, 2000) if the child was 6–7 years old. Secondary caregivers completed the Caregiver–Teacher Report Form (C–TRF; Achenbach & Rescorla, 2001) if the child was 3–5 years old or the Teacher's Report Form (TRF; Achenbach & Rescorla, 2001) if the child was 6–7 years old. The ASEBA scales are empirically derived, widely used, and have shown strong reliability (internal consistency, test–retest reliability, and interrater reliability) and validity (content, construct, and criterion-related validity) in large and diverse samples in the U.S. (Sattler, 2014).

Items on the CBCL 1.5–5 and C–TRF were categorized into seven syndrome scales: Emotionally Reactive, Anxious/Depressed, Somatic Complaints, Withdrawn, Sleep Problems (CBCL 1.5–5 only), Attention Problems, and Aggressive Behavior. Items on the CBCL 6–18 and TRF were categorized into eight syndrome scales: Anxious/Depressed, Somatic Complaints, Withdrawn/Depressed, Social Problems, Thought Problems, Attention Problems, Rule Breaking Behavior, and Aggressive Behavior. Subscales were further categorized into two higher-order factors: internalizing and externalizing. Scores on the Externalizing scale were used. The Externalizing scale consisted of the Attention Problems and Aggressive Behavior syndrome scales for the CBCL 1.5–5 (24 items) and C–TRF (34 items) and the Rule-breaking and Aggressive behavior syndrome scales for the CBCL 6–18 (35 items) and TRF (32 items). To account for missing responses in the sum score, scores were averaged across items and then multiplied by the number of items. As with self-regulation measures, externalizing problem scores were then converted to a proportion of the maximum possible score to put scores from different ASEBA measures onto a metric with the same possible range. Higher scores reflected more externalizing problems. Mothers' ratings on the Externalizing scale were associated with ratings by fathers ($r[113] = .60$, $p < .001$) and secondary caregivers ($r[109] = .45$, $p < .001$). Fathers' ratings were associated with ratings by secondary caregivers ($r[76] = .53$, $p < .001$). Age and sex norm-referenced *T*-scores had a mean of 46.51 (*SD* = 9.58).

Covariates. Covariates included the child's sex, age, grade, intelligence, and the family's socioeconomic status (SES).

Child's grade in school. The child's grade in school was coded as follows: 0 = not yet in kindergarten; 1 = kindergarten (or summer after kindergarten); 2 = first grade (or summer after 1st grade); 3 = second grade (or summer after second grade); 4 = third grade (or summer after third grade); 5 = fourth grade (or summer after fourth grade).

Child's intelligence. The child's intelligence was assessed using the age norm-referenced standard score for general cognitive ability on the Differential Abilities Scales-II (DAS; Elliott, 2007; Elliott et al., 2018). Parents completed the lower-level battery of the DAS if the child was between age 36 and 44 months or the upper-level battery if the child was between age 45 and 90 months. The lower-level battery consisted of four subtests: Naming Vocabulary, Pattern Construction, Verbal Comprehension, and Picture Similarities. The upper-level battery consisted of the same four subtests plus an additional two subtests: Matrices and Copying. Intelligence was assessed via the General Cognitive Ability standard score, which encompasses performance across all subtests at a given age. Higher scores reflected greater intelligence.

Family's socioeconomic status. The family's socioeconomic status (SES) was assessed based on the Four Factor Index of Social Status (Hollingshead, 1975). Parents reported on their occupation and education. Higher scores reflected higher socioeconomic status levels. Hollingshead scores suggested a sample with some variation in SES, but with a solid middle-class core.

Supplementary Appendix S4. Developmental Scaling Approach

We used developmental scaling to link scores from the different measures across ages onto the same scale. In this way, we could make meaningful comparisons of scores from different measures across ages and estimate accurate trajectories of children's self-regulation growth. To perform developmental scaling, we used a two-parameter Bayesian longitudinal item response model in a mixed modeling item response theory (IRT) framework. Such a model allows us to simultaneously account for heterotypic continuity of self-regulation using different measures across time and to model children' self-regulation trajectories. Given the numerous measures assessed, the many items, and the varying number of items per measure, we used measure-level (POM) scores (rather than item- and trial-level scores) as the "items" in the item-response model. The model linked scores from measures across all ages in the same model, known as concurrent calibration. Concurrent calibration accounts for within-person dependence of scores across time and results in more precise and stable estimates than two-stage calibration in which separate models are fit (Kolen & Brennan, 2014; McArdle et al., 2009). The two-parameter item response model estimates two parameters: easiness (ξ; the inverse of difficulty) and discrimination (α). The item's easiness parameter is the expected score on an item at a given level of the construct (Bürkner, 2020). The item's discrimination parameter is how strongly the item is associated with the construct. In our study, easiness and discrimination provide information about the functioning and usefulness of each measure—and the whole measurement scheme—at a given age. A two-parameter logistic IRT model takes the following form:

$$P\left(y_{ij} = 1 \middle| \theta_j, \alpha_i, \xi_i\right) = \frac{e^{\alpha_i\left(\theta_j + \xi_i\right)}}{1 + e^{\alpha_i\left(\theta_j + \xi_i\right)}} \qquad (1)$$

where y_{ij} is score for person j on item i, theta (θ_j) is the level on the construct for person j, xi (ξ_i) is the easiness parameter for item i, and alpha (α_i) is the discrimination parameter for item i.

In the present study, the self-regulation scores were continuous proportion scores that ranged from $p_{ij} = 0$–1. Because some scores were zero or one (especially one; see Supplementary Figure S2), we used a zero-one-inflated beta distribution for the outcome variable (Ospina & Ferrari, 2012). A traditional beta distribution is a continuous probability distribution that does not allow zeros or ones. A zero-one inflated beta distribution is a mixed continuous-discrete probability distribution, which includes a continuous beta distribution (to capture the continuous distribution of proportion scores) and a Bernoulli distribution (to capture zeros and ones). A zero-one-inflated beta response distribution takes the following form:

$$f(p_{ij}) = \begin{cases} \pi_{ij} & \text{if } p_{ij} = 0 \\ (1 - \pi_{ij})\gamma_{ij} & \text{if } p_{ij} = 1 \\ (1 - \pi_{ij})(1 - \gamma_{ij})\text{Beta}(a_{ij}, b_{ij}) & \text{if } p_{ij} \in (0, 1) \end{cases} \tag{2}$$

where pi (π_{ij}) is the probability of $p_{ij} = 0$, gamma (γ_{ij}) is the conditional probability of $p_{ij} = 1$ given that $p_{ij} \neq 0$, and a_{ij} and b_{ij} are the shape parameters of the Beta distribution when $p_{ij} \in (0, 1)$.

The next step in the Bayesian hierarchical model is to put distributions on each of the parameters in Equation 2. We estimated both π_{ij} and γ_{ij} using a logistic mixed model with fixed effects for the child's age and sex along with a random intercept for subject.

Nesting:

Level 1: i = item (i.e., measure of self-regulation)

Level 2: j = person

$$\text{logit}(\pi_{ij}) = \beta_{0\pi} + \beta_{1\pi}\text{age}_{ij} + \beta_{2\pi}\text{sex}_j + b_{0\pi j} \tag{3}$$

$$\text{logit}(\gamma_{ij}) = \beta_{0\gamma} + \beta_{1\gamma}\text{age}_{ij} + \beta_{2\gamma}\text{sex}_j + b_{0\gamma j}$$

The a_{ij} and b_{ij} parameters in the beta distribution are re-written into the mean $\mu_{ij} = \frac{a_{ij}}{a_{ij}+b_{ij}}$ and the variance $v_{ij} = \frac{a_{ij} \times b_{ij}}{\left(a_{ij}+b_{ij}\right)^2 \left(a_{ij}+b_{ij}+1\right)}$. The mean μ_{ij} is a percentage that is given the IRT form of Equation 1 and specified in more detail below in Equation 4.

$$\text{logit}\left(\mu_{ij}\right) = e^{\log(\alpha_i)} \times \eta_{ij} \tag{4}$$

$$\eta_{ij} = \theta_j + \xi_i$$

where mu (μ_{ij}) is the self-regulation score for person j on item i, alpha (α_i) is the discrimination parameter for item i, eta (η_{ij}) is the sum of the person's level on the construct (θ_j) for person j and the item's easiness (ξ_i) for item i.

In addition, we took the log of the variance and used a linear mixed model on v_{ij} using an intercept-only log-linear mixed model with a population intercept and a random intercept for subject.

$$\log\left(v_{ij}\right) = \beta_{0\gamma} + b_{0\gamma j} \tag{5}$$

We performed the developmental scaling, estimation of growth curves, and tests of differential item (measure) functioning (DIF) in the same model. A given child had up to four time points. Thus, a quadratic was the most complex polynomial of nonlinear growth we could estimate for children's trajectories that still allow measurement error. Because of prior work demonstrating that growth in self-regulation is non-linear, such that children showed faster growth in preschool than elementary school (Montroy et al., 2016), we modeled children's growth in self-regulation with a quadratic term. We modeled random intercepts and random linear and quadratic slopes to allow each child to differ in their starting point, form of growth, and curvature. Age in years was centered to set the intercepts at age 3. We included the child's sex (female = 1, male = 0) as a predictor of the intercepts and slopes.

To examine DIF across ages, we estimated a random intercept and slope for measure (in addition to the terms described above) to allow the item parameters for each measure to differ by age. This allowed us to examine the extent to which the measures changed across development in easiness and discrimination. Specifically, we estimated the model according to the following equation:

$$\eta_{itj} = \theta_j + \xi_{it} + \beta_{01}(\text{age}_{tj}) + \beta_{02}(\text{age}_{tj}^2) + \beta_{03}(\text{sex}_j) + \beta_{04}(\text{sex}_j \times \text{age}_{tj}) + \beta_{05}(\text{sex}_j \times \text{age}_{tj}^2) + b_i + b_{01i}(\text{age}_{tj}) + b_p + b_{11j}(\text{age}_{tj}) + b_{12j}(\text{age}_{tj}^2) \quad (6)$$

$$\log(\alpha_{it}) = \beta_0 + \beta_{11}(\text{age}_{tj}) + \beta_{12}(\text{age}_{tj}^2) + b_{0i} + b_{11i}(\text{age}_{tj})$$

where alpha (α_{it}) is the discrimination parameter for item i at time t, as predicted by a random intercept for the item, fixed effects of linear and quadratic age, and a random effect of linear age; eta (η_{itj}) is the sum of the person's level on the construct (θ_j) for person j and the item's easiness (ξ_{it}) for item i at time t, in addition to random intercepts for subject, fixed and random effects of linear and quadratic age, and the child's sex (both as a main effect and interaction with linear and quadratic age). Thus, the model handles the longitudinal dependency of data within subject.

In a Bayesian model, the final step is to specify prior distributions for all remaining parameters in the model. We kept the default priors used in the brms package (Bürkner, 2017), which uses vague but proper priors. The priors were logistic (mean 0, scale parameter 1) for the intercept of the probability of having a score of 0 or 1 (zero-one inflation; zoi) and the conditional probability of having a score of 1 given the score is either 0 or 1 (conditional one-inflation; coi). The intercept for precision (phi; i.e., 1/variance) and all standard deviation parameters were given a half *t*-distribution prior with 3 degrees of freedom, mean 0, and scale parameter 2.5.

Our model had no missing data in the predictors (age and sex); missingness was only in the outcome (scores on self-regulation measures). Mixed models handle missing data in the outcomes. Mixed models provide valid inferences if the data are missing at random or completely at random. Because much of our missingness was due to COVID-19, and we observed limited patterns of systematic missingness as a function of demographics, predictors, or outcomes with small effect sizes, we felt this modeling approach was appropriate. Moreover, researchers have argued against using multiple imputation in longitudinal designs that use mixed models because multiple imputation can lead to unstable estimates (Twisk et al., 2013).

Developmentally scaled self-regulation factor scores were estimated from the posterior distribution by averaging model-predicted posterior samples across chains and iterations, within combinations of child-by-measurement occasion. This allowed each child to have a different factor score at each of their measurement occasions.

We fit the Bayesian longitudinal mixed model using the brms package 2.16.3 (Bürkner, 2017) in R, which uses the RStan 2.21.3 (Stan Development Team, 2020a) interface to Stan 2.21.0 (Stan Development Team, 2020b) for Bayesian modeling. The model included eight chains and 10,000 iterations.

Supplementary Appendix S5. Longitudinal Measurement Invariance and Sensitivity Analyses.

Testing measurement invariance is important because identifying longitudinal measurement non-invariance can provide evidence of heterotypic continuity (Edwards & Wirth, 2009; Nesselroade & Estabrook, 2009; Widaman et al., 2010). However, establishing longitudinal measurement invariance should not be considered necessary when the construct shows heterotypic continuity, because the factor structure of the construct, by definition, changes with development (Knight & Zerr, 2010; Petersen et al., 2020) and models with failed longitudinal measurement invariance can still yield valid inferences when the construct shows heterotypic continuity (Edwards & Wirth, 2012; Lai, 2021).

We examined longitudinal measurement invariance. Our original model showed changes in item easiness and discrimination across ages, consistent with heterotypic continuity. Changes in item easiness and discrimination are depicted in Supplementary Figure S3. Four measures became easier—relative to the same ability—with age: Fish/Sharks, Gift Delay, Snack Delay, and mothers' ratings on the BRIEF–P. All measures except Fish/Sharks and Simon Says showed decreases in discrimination with age, consistent with heterotypic continuity. However, effect sizes of non-invariance were small, and measures remained strongly discriminating across ages.

When the construct shows heterotypic continuity, removing non-invariant items can be problematic because it can reduce content validity (Knight & Zerr, 2010), the reliability of measurement, and the ability to detect individual differences (i.e., person separation). And removing non-invariant measures would leave us essentially no measures (two) to examine the growth of self-regulation, which provides further evidence on the nature of the developing construct. Prior research has examined self-regulation growth with only one or a few measures in each study. Given the complex, changing, and multi-faceted nature of the construct, using only one or two measures is an unsatisfactory solution to examine children’s growth in self-regulation. One or a few measures would likely not capture the changing manifestation of the construct. Instead, consistent with our approach, *resolving* differential item functioning by allowing some item parameters to differ across ages has very little effect on reliability and person separation, and is generally recommended in favor of removing non-invariant items (Hagquist, 2019; Hagquist & Andrich, 2017). Resolving, rather than removing, measures is especially relevant for our study because the measures were associated with the construct at each age they were assessed, and the measures were still strongly associated with the latent construct even after controlling for the child’s age in exploratory factor analysis. Thus, we took several steps to minimize and evaluate the impacts of longitudinal measurement invariance.

First, we resolved differential item functioning (i.e., longitudinal non-invariance) of one of the measures (the BRIEF) by allowing it to have different item parameters at early ages (ages 3–5; BRIEF–P) compared to later ages (ages 6–7; BRIEF–2).

Second, we fit an additional model with longitudinal invariance constraints. In frequentist approaches, measurement invariance is often tested in a four-step approach (configural, scalar, metric, and residual invariance). In a Bayesian approach, however, a conditional logic structure is preferred to sequential model testing because the models provide the information necessary to evaluate the extent of any non-invariance. The advantage of the conditional formulation of Bayesian mixed models is well described in many excellent textbooks (Cowles, 2013; Gelman et al., 2013; Kruschke, 2014) and the documentation for the R package brms (Bürkner, 2017). Wikle et al. (2019, p. 10) wrote: “If most of the complex dependencies in the data are due to the underlying process of interest, then one should model the distribution of the data conditioned on that process (data model), followed by a model of the process' behavior and its uncertainties (process model)." Therefore, within a Bayesian mixed modeling framework, approximate measurement invariance is a process of interest which can be used to account for small instances of non-invariance (van de Schoot et al., 2013, 2015), as exist in the present study. Approximate measurement invariance involves setting narrow priors on the invariance parameters rather than fixing invariance parameters to zero (van de Schoot et al., 2013). Approximate measurement invariance is more accurate than full or partial measurement invariance for estimating true latent mean differences when there are many small differences in the intercepts and factor loadings across groups (Cieciuch et al., 2014; van de Schoot et al., 2013). Thus, as a sensitivity analysis in the present study, we also fit a model that imposes approximate longitudinal measurement invariance. In this model, we set the slopes of the discrimination parameters to be close to zero, by setting the prior of the discrimination parameter to have a normal distribution with a mean of zero and a small standard deviation of 0.05. We also set the prior of the standard deviation of the random effect of task on the association between age and discrimination to be small (normal distribution with mean = 0, *SD* = 0.01) so that tasks were restricted to be similar in their change of discrimination parameter (i.e., near zero). Using our modeling approach, we are unable to fix the measures' easiness to be the same across ages because the eta parameter is the sum of the person's level on the construct and the item easiness (Bürkner, 2020), and restricting people's levels on the construct to be the same across ages would thus prevent us from being able to detect growth. This approximate measurement invariance approach successfully constrained the factor loadings to be near-zero, and no factor loadings showed significant differences across ages in this model.

Third, as an additional test, we also conducted a sensitivity test by fitting a Bayesian model that excluded scores for a given measure at ages when the proportion of maximum score on that measure could reflect ceiling effects (i.e., mean proportion score > .90). No measures appeared to show mean-level floor effects.

Findings were substantially similar when examining the model with approximate longitudinal variance imposed, and when examining the model with potential mean-level ceiling effects. Both models yielded similar trajectories of self-regulation: i.e., rapid growth in self-regulation from ages 3–6, after which growth slowed and leveled off. In addition, criterion-related tests yielded similar results. Self-regulation was negatively associated with externalizing problems, controlling for age. Self-regulation was positively associated with math and reading skills when controlling for age, grade, and SES. Self-regulation was positively associated with math (but not reading) skills when adding a control for intelligence.

Supplementary Appendix S6. Power Analyses.

We conducted a power analysis to determine our probability to detect the expected effect sizes of self-regulation in predicting externalizing problems and school readiness. The smallest association we would be able to detect with a power of .8 is *r* = |.18|. Based on recent meta-analyses, the effect size of self-regulation in association with externalizing problems is *r* = -.11 (Berger & Buttelmann, 2021) and with school readiness is *r* = .37 (Robson et al., 2020). Based on our sample size, we would have power (i.e., probability to detect associations) of .39 and .99 to detect the association of self-regulation with externalizing problems and school readiness, respectively. Thus, we are well-powered to detect an association of self-regulation with school readiness. By contrast, we are somewhat under-powered to detect an association of self-regulation with externalizing problems.

References

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA Preschool Forms and Profiles: An integrated system of multi-informant assessment.* Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles.* Burlington, VT: University of Vermont, Department of Psychiatry.

Berger, A., Alyagon, U., Hadaya, H., Atzaba-Poria, N., & Auerbach, J. G. (2013). Response inhibition in preschoolers at familial risk for Attention Deficit Hyperactivity Disorder: A behavioral and electrophysiological stop-signal study. *Child Development*, *84*(5), 1616–1632. https://doi.org/10.1111/cdev.12072

Berger, P., & Buttelmann, D. (2021). A meta-analytic approach to the association between inhibitory control and parent-reported behavioral adjustment in typically-developing children: Differentiating externalizing and internalizing behavior problems. *Developmental Science*, e13141. https://doi.org/10.1111/desc.13141

Bürkner, P.-C. (2017). brms: An R Package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2020). Bayesian item response modeling in R with brms and Stan. *ArXiv:1905.09501*. https://doi.org/10.48550/arXiv.1905.09501

Carlson, S. M., Davis, A. C., & Leach, J. G. (2005). Less is More: Executive function and symbolic representation in preschool children. *Psychological Science*, *16*(8), 609–616. https://doi.org/10.1111/j.1467-9280.2005.01583.x

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, *72*(4), 1032–1053. https://doi.org/10.1111/1467-8624.00333

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (2014). Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: A cross-country illustration with a scale to measure 19 human values. *Frontiers in Psychology*, *5*, 982. https://doi.org/10.3389/fpsyg.2014.00982

Cowles, M. K. (2013). *Applied bayesian statistics: With R and OpenBUGS examples*. Springer.

Duckworth, A. L., Tsukayama, E., & Kirby, T. A. (2013). Is it really self-control? Examining the predictive power of the delay of gratification task. *Personality and Social Psychology Bulletin*, *39*(7), 843–855. https://doi.org/10.1177/0146167213482589

Edwards, M. C., & Wirth, R. J. (2009). Measurement and the study of change. *Research in Human Development*, *6*(2–3), 74–96. https://doi.org/10.1080/15427600902911163

Edwards, M. C., & Wirth, R. J. (2012). Valid measurement without factorial invariance: A longitudinal example. In J. R. Harring & G. R. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences* (pp. 289–311). IAP Information Age Publishing.

Eisenberg, N., Edwards, A., Spinrad, T. L., Sallquist, J., Eggum, N. D., & Reiser, M. (2013). Are effortful and reactive control unique constructs in young children? *Developmental Psychology*, *49*(11), 2082–2094. https://doi.org/10.1037/a0031745

Elliott, C. D. (2007). *Differential Ability Scales (2nd ed.).* San Antonio, TX: Harcourt Assessment.

Elliott, C. D., Salerno, J. D., Dumont, R., & Willis, J. O. (2018). The Differential Ability Scales—Second Edition. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 360–382). The Guilford Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd Edition).

Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: Performance of children 3 1/2-7 years old on a Stroop-like day-night test. *Cognition*, *53*(2), 129–153. https://doi.org/10.1016/0010-0277(94)90068-x

Gioia, G. A., Andrwes, K., & Isquith, P. K. (1996). *Behavior Rating Inventory of Executive Function: Preschool version (BRIEF-P) [Measurement Instrument].* Odessa, FL: Psychological Assessment Resources.

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2015). *Behavior Rating Inventory of Executive Function: Second edition (BRIEF2) [Measurement Instrument].* Odessa, FL: Psychological Assessment Resources.

Goldsmith, H. H., Reilly, J., Lemery, K. S., Longley, S., & Prescott, A. (1999). *The laboratory temperament assessment battery: Preschool version.* Unpublished Manuscript.

Hagquist, C. (2019). Explaining differential item functioning focusing on the crucial role of external information – an example from the measurement of adolescent mental health. *BMC Medical Research Methodology*, *19*(1), 185. https://doi.org/10.1186/s12874-019-0828-3

Hagquist, C., & Andrich, D. (2017). Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health and Quality of Life Outcomes*, *15*(1), 181. https://doi.org/10.1186/s12955-017-0755-0

Hollingshead, A. B. (1975). *Four-factor index of social status.* Unpublished manuscript, Yale University, New Haven, CT.

Klenberg, L., Korkman, M., & Lahti-Nuuttila, P. (2001). Differential development of attention and executive functions in 3- to 12-year-old Finnish children. *Developmental Neuropsychology*, *20*(1), 407–428. https://doi.org/10.1207/S15326942DN2001_6

Knight, G. P., & Zerr, A. A. (2010). Informed theory and measurement equivalence in child development research. *Child Development Perspectives*, *4*(1), 25–30. https://doi.org/10.1111/j.1750-8606.2009.00112.x

Kochanska, G., Murray, K., & Coy, K. C. (1997). Inhibitory control as a contributor to conscience in childhood: From toddler to early school age. *Child Development*, *68*(2), 263–277. https://doi.org/10.2307/1131849

Kochanska, G., Murray, K., Jacques, T. Y., Koenig, A. L., & Vandegeest, K. A. (1996). Inhibitory control in young children and its role in emerging internalization. *Child Development*, *67*(2), 490–507.

Kochanska, G., Murray, K. T., & Harlan, E. T. (2000). Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. *Developmental Psychology*, *36*(2), 220–232. https://doi.org/10.1037/0012-1649.36.2.220

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. Springer Science & Business Media.

Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Lai, M. H. C. (2021). Adjusting for measurement noninvariance with alignment in growth modeling. *Multivariate Behavioral Research*, *0*(0), 1–18. https://doi.org/10.1080/00273171.2021.1941730

Luria, A. R., Pribram, K. H., & Homskaya, E. D. (1964). An experimental analysis of the behavioral disturbance produced by a left frontal arachnoidal endothelioma (meningioma). *Neuropsychologia*, *2*(4), 257–280. https://doi.org/10.1016/0028-3932(64)90034-X

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods*, *14*(2), 126–149. https://doi.org/10.1037/a0015857

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review*, *106*(1), 3–19. https://doi.org/10.1037/0033-295X.106.1.3

Mischel, W., Ebbesen, E. B., & Raskoff Zeiss, A. (1972). Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology*, *21*(2), 204–218. https://doi.org/10.1037/h0032198

Montroy, J. J., Bowles, R. P., Skibbe, L. E., McClelland, M. M., & Morrison, F. J. (2016). The development of self-regulation across early childhood. *Developmental Psychology*, *52*(11), 1744–1762. https://doi.org/10.1037/dev0000159

Nesselroade, J. R., & Estabrook, R. (2009). Factor invariance, measurement, and studying development over the lifespan. In H. C. Bosworth & C. Hertzog (Eds.), *Aging and* *cognition: Research methodologies and empirical advances* (pp. 39–52). American Psychological Association. https://doi.org/10.1037/11882-002

Neuenschwander, R., & Blair, C. (2017). Zooming in on children's behavior during delay of gratification: Disentangling impulsigenic and volitional processes underlying self-regulation. *Journal of Experimental Child Psychology*, *154*, 46–63. https://doi.org/10.1016/j.jecp.2016.09.007

Ospina, R., & Ferrari, S. L. P. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, *56*(6), 1609–1623. https://doi.org/10.1016/j.csda.2011.10.005

Peake, P. K. (2017). Delay of gratification: Explorations of how and why children wait and its linkages to outcomes over the life course. In J. R. Stevens (Ed.), *Impulsivity* (pp. 7–60). Springer, Cham. https://doi.org/10.1007/978-3-319-51721-6_2

Petersen, I. T., Choe, D. E., & LeBeau, B. (2020). Studying a moving target in development: The challenge and opportunity of heterotypic continuity. *Developmental Review*, *58*, 100935. https://doi.org/10.1016/j.dr.2020.100935

Putnam, S. P., & Rothbart, M. K. (2006). Development of short and very short forms of the Children's Behavior Questionnaire. *Journal of Personality Assessment*, *87*(1), 102–112. https://doi.org/10.1207/s15327752jpa8701_09

Razza, R. A., & Raymond, K. (2013). Associations among maternal behavior, delay of gratification, and school readiness across the early childhood years. *Social Development*, *22*(1), 180–196. https://doi.org/10.1111/j.1467-9507.2012.00665.x

Robson, D. A., Allen, M. S., & Howard, S. J. (2020). Self-regulation in childhood as a predictor of future outcomes: A meta-analytic review. *Psychological Bulletin*, *146*(4), 324–354. https://doi.org/10.1037/bul0000227

Sattler, J. M. (2014). *Foundations of behavioral, social, and clinical assessment of children (6th ed.)*. Jerome M. Sattler, Publisher, Inc.

Schneider, W., Eschman, A., & Zuccolotto, A. (2012). *E-Prime Reference Guide*. Pittsburgh: Psychology Software Tools, Inc.

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). *Woodcock-Johnson IV Tests of Cognitive Abilities.* Rolling Meadows, IL: Riverside.

Schrank, F. A., Wendling, B. J., Flanagan, D. P., & McDonough, E. M. (2018). The Woodcock–Johnson IV Tests of Early Cognitive and Academic Development. In *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 283–301). The Guilford Press.

Stan Development Team. (2020b). *RStan: The R interface to Stan. R package version 2.21.0.* http://mc-stan.org/

Stan Development Team. (2020a). *RStan: The R interface to Stan. R package version 2.21.3*. http://mc-stan.org/

Strommen, E. A. (1973). Verbal self-regulation in a children's game: Impulsive errors on "Simon says." *Child Development*, *44*(4), 849–853. https://doi.org/10.2307/1127737

Teglasi, H., Schussler, L., Gifford, K., Annotti, L. A., Sanders, C., & Liu, H. (2015). Child behavior questionnaire–short form for teachers: Informant correspondences and divergences. *Assessment*, *22*(6), 730–748. https://doi.org/10.1177/1073191114562828

Twisk, J., de Boer, M., de Vente, W., & Heymans, M. (2013). Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis.

Journal of Clinical Epidemiology, *66*(9), 1022–1028. https://doi.org/10.1016/j.jclinepi.2013.03.017

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthen, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*, 770. https://doi.org/10.3389/fpsyg.2013.00770

van de Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement invariance. *Frontiers in Psychology, 6*, 1064. https://doi.org/10.3389/fpsyg.2015.01064

Verbruggen, F., Aron, A. R., Band, G. P., Beste, C., Bissett, P. G., Brockett, A. T., Brown, J. W., Chamberlain, S. R., Chambers, C. D., Colonius, H., Colzato, L. S., Corneil, B. D., Coxon, J. P., Dupuis, A., Eagle, D. M., Garavan, H., Greenhouse, I., Heathcote, A., Huster, R. J., … Boehler, C. N. (2019). A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. *ELife*, *8*, e46323. https://doi.org/10.7554/eLife.46323

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10–18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

Wiebe, S. A., Sheffield, T. D., & Espy, K. A. (2012). Separating the fish from the sharks: A longitudinal study of preschool response inhibition. *Child Development*, *83*(4), 1245–1261. https://doi.org/10.1111/j.1467-8624.2012.01765.x

Wikle, C. K., Zammit-Mangion, A., & Cressie, N. (2019). *Spatio-Temporal statistics with R* (1st edition). Chapman and Hall/CRC.

Supplementary Table S1

Percent of ID-wave instances with scores on each measure (out of eligible instances)

Measure	Percent of ID-Wave instances with Scores
Developmentally Scaled Self-Regulation	78.98
Bear/Dragon	61.69
BRIEF: Mother	73.90
BRIEF: Father	37.29
CBQ: Mother	74.92
CBQ: Father	40.00
CBQ: Secondary Caregiver	36.95
Day/Night	60.00
Fish Sharks	51.19
Gift Delay	61.36
Grass/Snow	58.31
Hand Game	56.95
Knock Tap	61.02
Less is More	53.56
Peg Tapping	59.66
Self-Imposed Waiting Task	60.34
Shape Stroop	63.05
Simon Says	58.98
Snack Delay	62.37
Stop Signal Task	36.95
Token Sort	60.00
CBCL: Mother	75.59
CBCL: Father	42.11
(C–)TRF	38.98
Reading skills	63.05
Math skills	63.05

Note. “BRIEF” = Behavior Rating Inventory of Executive Function; “CBQ” = Children’s Behavior Questionnaire; “CBCL” = Child Behavior Checklist; “C–TRF” = Caregiver–Teacher Report Form; “TRF” = Teacher’s Report Form

Supplementary Table S2

Self-Regulation Bivariate Correlations and Descriptive Statistics

		BRIEF:	BRIEF:	CBQ:	CBQ:	CBQ:															
	BD	M	F	M	F	S	DN	DoG	FS	GD	GS	HG	KT	LiM	PT	SH	SI	SD	SST	TS	SR
BD	–																				
BRIEF:M	.07	–																			
BRIEF:F	.14	.35***	–																		
CBQ:M	.25***	.38***	.37***	–																	
CBQ:F	.35***	.33***	.57***	.51***	–																
CBQ:S	.06	.27*	.28*	.38***	.43***	–															
DN	.58***	.10	.15	.25***	.37***	.06	–														
DoG	.58***	.01	.08	.23***	.36***	.14	.44***	–													
FS	.48***	.15†	.26*	.23*	.30*	.04	.35***	.44***	–												
GD	.37***	.02	-.10	.19*	.16	.05	.33***	.35***	.32***	–											
GS	.69***	.16*	.27*	.33***	.34***	.15	.68***	.51***	.43***	.39***	–										
HG	.62***	.10	.32*	.18*	.33***	.02	.50***	.47***	.54***	.20*	.59***	–									
KT	.72***	.14†	.23*	.29***	.44***	.19†	.66***	.53***	.59***	.39***	.72***	.72***	–								
LiM	.51***	.03	.16	.24***	.29*	.29*	.41***	.39***	.23*	.24***	.42***	.34***	.43***	–							
PT	.66***	.16*	.30*	.25***	.34***	.05	.62***	.51***	.50***	.35***	.73***	.63***	.75***	.46***	–						
SH	.39***	.07	.19†	.21*	.38***	.11	.23***	.31***	.55***	.39***	.32***	.35***	.43***	.19*	.31***	–					
SI	.59***	-.06	.12	.14†	.29*	.06	.57***	.43***	.43***	.34***	.55***	.43***	.53***	.38***	.59***	.26***	–				
SD	.41***	.06	.13	.25***	.33***	.18†	.20*	.35***	.52***	.34***	.28***	.36***	.41***	.18*	.24***	.50***	.23***	–			
SST	.25*	-.02	.23	.27*	.26†	.03	.28***	.29***	.34***	.11	.35***	.13	.24*	.11	.32***	.10	.18†	.20*	–		
TS	.42***	.07	.13	.20*	.30***	.10	.27***	.35***	.25***	.28***	.32***	.47***	.48***	.38***	.48***	.27***	.26***	.32***	.10	–	
SR	.81***	.18*	.29***	.38***	.47***	.21*	.75***	.64***	.60***	.45***	.82***	.73***	.84***	.57***	.82***	.44***	.75***	.40***	.36***	.51***	–
n	182	218	110	221	118	109	177	178	151	181	172	168	180	158	176	186	174	184	109	177	233
M	0.70	0.74	0.72	0.67	0.63	0.66	0.65	0.65	0.88	0.77	0.70	0.73	0.71	0.76	0.63	0.92	0.45	0.92	0.74	0.43	0.76
SD	0.32	0.16	0.15	0.09	0.08	0.12	0.34	0.44	0.12	0.18	0.37	0.31	0.34	0.27	0.34	0.21	0.29	0.15	0.18	0.24	0.60

Note. “BD” = Bear/Dragon; “BRIEF” = Behavior Rating Inventory of Executive Function; “CBQ” = Children’s Behavior Questionnaire”; “:M” = Mother-report; “:F” = Father-report; “:S” = Secondary caregiver-report; “DN” = Day/Night; “DoG” = Delay of gratification (self-imposed waiting task); “FS” = Fish/Sharks; “GD” = Gift Delay; “GS” = Grass/Snow; “HG” = Hand Game; “KT” = Knock/Tap; “LiM” = Less is More; “PT” = Peg Tapping; “SH” = Shape Stroop; “SI” = Simon Says; “SD” = Snack Delay; “SST” = stop-signal task; “TS” = Token Sort; “SR” = developmentally scaled self-regulation score

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

Supplementary Table S3

ID-Wave Bivariate Correlations and Descriptive Statistics

	sex	age	grade	SES	SR	reading	math	intelligence
sex	–							
age	.10	–						
grade	.02	.80***	–					
SES	-.08	.23***	.17*	–				
SR	.23***	.63***	.31***	.30***	–			
reading	.13†	.73***	.64***	.27***	.65***	–		
math	.13†	.82***	.56***	.35***	.86***	.76***	–	
intelligence	.14†	-.01	-.01	.21*	.37***	.23***	.39***	–
n	432	235	194	428	233	186	186	156
M	0.47	4.82	0.43	51.10	0.76	12.17	12.67	109.10
SD	0.50	1.22	0.80	10.88	0.60	12.84	6.68	16.22

Note. The last available socioeconomic status (SES) scores were carried forward to fill in missing values for a given participant.

“SES” = socioeconomic status; “SR” = developmentally scaled self-regulation score

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

Supplementary Table S4

ID-Wave-Rater Bivariate Correlations and Descriptive Statistics

	sex	age	SES	SR	EXT	BRIEF	CBQ
sex	–						
age	.13*	–					
SES	-.02	.21***	–				
SR	.28***	.65***	.33***	–			
EXT	-.14***	-.32***	-.19***	-.28***	–		
BRIEF	.11†	-.02	.12*	.21***	-.52***	–	
CBQ	.26***	.17***	.16***	.35***	-.39***	.44***	–
n	668	471	663	469	469	332	453
M	0.48	4.72	52.62	0.74	0.15	0.73	0.66
SD	0.50	1.24	10.31	0.61	0.15	0.16	0.10

Note. The last available socioeconomic status (SES) scores were carried forward to fill in missing values for a given participant.

“SES” = socioeconomic status; “SR” = developmentally scaled self-regulation score; “EXT” = externalizing problems; “BRIEF” = Behavior Rating Inventory of Executive Function; “CBQ” = Children’s Behavior Questionnaire

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

Supplementary Table S5

Reliability of Measures

Measure	Inter-rater Reliability	Internal Consistency Reliability	Cross-time 9-Month Stability
Developmentally Scaled Self-Regulation	n/a	ω = .94	*r* = .68
Bear/Dragon	ICC[2,*k*] = .99	Mean reliability of all possible split halves: go: .83; no-go: .96	*r* = .48
BRIEF: Mother	n/a	α = .96 (BRIEF–P); α = .97 (BRIEF–2)	*r* = .62
BRIEF: Father	n/a	α = .95 (BRIEF–P); α = .97 (BRIEF–2)	*r* = .75
CBQ: Mother	n/a	ω = .87	*r* = .81
CBQ: Father	n/a	ω = .85	*r* = .72
CBQ: Secondary Caregiver	n/a	ω = .87	*r* = .56
Day/Night	ICC[2,*k*] = .99	Mean reliability of all possible split halves: .94	*r* = .51
Fish/Sharks	n/a	Mean reliability of 1,000,000 split halves: go: .92; no-go: .94	*r* = .36
Gift Delay	ICC[2,*k*] = .99	n/a	*r* = .29
Grass/Snow	ICC[2,*k*] = .99	Mean reliability of all possible split halves: .96	*r* = .44
Hand Game	ICC[2,*k*] = .99	Mean reliability of all possible split halves: .93	*r* = .33
Knock Tap	ICC[2,*k*] = .99	Mean reliability of all possible split halves: .95	*r* = .53
Less is More	ICC[2,*k*] = .99	Mean reliability of all possible split halves: .90	*r* = .39
Peg Tapping	ICC[2,*k*] = .99	Mean reliability of all possible split halves: .93	*r* = .50
Self-Imposed Waiting Task	ICC[2,*k*] = .99	n/a	*r* = .37
Shape Stroop	ICC[2,*k*] = .99	Mean reliability of all possible split halves: .84	*r* = .54
Simon Says	ICC[2,*k*] = .99	Mean reliability of all possible split halves: go: .75; no-go: .94	*r* = .69
Snack Delay	ICC[2,*k*] = .99	Mean reliability of all possible split halves: .96	*r* = .24
Stop-Signal Task	n/a	n/a	*r* = .64
Token Sort	n/a	n/a	*r* = .56
CBCL: Mother	n/a	ω_C = .98 (CBCL 1.5–5); α = .84 (CBCL 6–18)	*r* = .62
CBCL: Father	n/a	α = .89 (CBCL 1.5–5); α = .86 (CBCL 6–18)	*r* = .78
(C–)TRF	n/a	α = .95 (C–TRF); α = .86 (TRF)	*r* = .72
Reading skills	n/a	*	*r* = .88

Math skills	n/a	*	*r* = .88

Note. Cronbach's alpha was calculated only if omega (ω) was unable to be calculated (due to non-convergence) or if the estimate was above 1.0. Split-half reliability was corrected for the number of items of the split halves relative to the number of items in the measure, using the Spearman–Brown prophecy formula. For Fish/Sharks, a random sample of 1,000,000 split halves was used for calculating the mean split-half reliability rather than all possible split halves due to the large number of trials and the lengthy time it would take to compute the mean reliability of all possible split halves.

ω_C = omega categorical (for items with fewer than 5 response categories); "BRIEF" = Behavior Rating Inventory of Executive Function; "CBQ" = Children's Behavior Questionnaire; "CBCL" = Child Behavior Checklist; "C–TRF" = Caregiver–Teacher Report Form; "TRF" = Teacher's Report Form

* Item-level scores were not entered to be examined

Supplementary Table S6

Duration to Complete Each Task (Minutes)

Measure	*M*	*SD*
Bear/Dragon	4.76	1.34
BRIEF	5.11	2.49
CBQ	23.99	11.72
CBQ–TSF	15.40	7.41
Day/Night	2.02	1.06
Fish/Sharks	*	*
Gift Delay	5.22	0.64
Grass/Snow	2.41	1.92
Hand Game	3.58	1.48
Knock Tap	3.61	1.85
Less is More	4.15	0.80
Peg Tapping	2.43	0.98
Self-Imposed Waiting Task	8.57	3.24
Shape Stroop	1.40	0.61
Simon Says	2.45	1.45
Snack Delay	4.01	1.46
Stop-Signal Task	*	*
Token Sort	*	*
CBCL	11.81	6.22
(C–)TRF	7.03	3.46
Reading skills	*	*
Math skills	*	*

“BRIEF” = Behavior Rating Inventory of Executive Function; “CBQ” = Children’s Behavior Questionnaire; “CBQ–TSF” = Children’s Behavior Questionnaire–Teacher Short Form; “CBCL” = Child Behavior Checklist; “C–TRF” = Caregiver–Teacher Report Form; “TRF” = Teacher’s Report Form

* Task times were not recorded.

Supplementary Table S7

ID-Wave Partial Correlations Controlling for Age

	sex	grade	SES	SR	reading	math	intelligence
sex	–						
grade	-.10	–					
SES	-.10	-.02	–				
SR	$.22^{***}$	$-.41^{***}$	$.20^{***}$	–			
reading	.08	$.13^{\dagger}$	$.16^{*}$	$.37^{***}$	–		
math	.09	$-.27^{***}$	$.28^{***}$	$.77^{***}$	$.42^{***}$	–	
intelligence	$.14^{\dagger}$	-.01	$.22^{*}$	$.48^{***}$	$.35^{***}$	$.68^{***}$	–

Note. “SES” = socioeconomic status; “SR” = developmentally scaled self-regulation score

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

Supplementary Table S8

Self-Regulation Partial Correlations Controlling for Age

		BRIEF:	BRIEF:	CBQ:	CBQ:	CBQ:															
	BD	M	F	M	F	S	DN	DoG	FS	GD	GS	HG	KT	LiM	PT	SH	SI	SD	SST	TS	SR
BD	–																				
BRIEF:M	.14†	–																			
BRIEF:F	.14	.36***	–																		
CBQ:M	.18*	.40***	.36***	–																	
CBQ:F	.24*	.36***	.58***	.49***	–																
CBQ:S	.00	.28***	.28*	.37***	.42***	–															
DN	.35***	.17*	.15	.19*	.28*	.01	–														
DoG	.39***	.05	.07	.16*	.27*	.11	.21*	–													
FS	.23***	.22*	.28*	.16†	.20†	-.02	.06	.22*	–												
GD	.19*	.05	-.13	.13†	.07	.02	.16*	.21*	.16†	–											
GS	.52***	.24***	.30*	.28***	.24*	.11	.53***	.30***	.19*	.24***	–										
HG	.49***	.15†	.33***	.11	.24*	-.03	.33***	.31***	.40***	.05	.45***	–									
KT	.55***	.23***	.26*	.23***	.37***	.17†	.47***	.31***	.39***	.23***	.58***	.63***	–								
LiM	.36***	.06	.15	.19*	.20†	.27*	.25***	.24***	.02	.11	.27***	.19*	.25***	–							
PT	.44***	.27***	.35***	.18*	.24*	-.02	.41***	.27***	.25***	.16*	.538***	.49***	.58***	.30***	–						
SH	.23*	.10	.19†	.16*	.32***	.09	.04	.15*	.46***	.29***	.16*	.23***	.29***	.06	.11	–					
SI	.27***	-.02	.12	.01	.14	-.01	.29***	.10	.10	.12	.26***	.18*	.17*	.16†	.26***	.01	–				
SD	.25***	.09	.12	.20*	.26*	.16	.00	.21*	.41***	.24***	.11	.25***	.26***	.04	.02	.43***	-.03	–			
SST	.11	.00	.22	.24*	.21*	.01	.16†	.18†	.24*	.01	.25*	.01	.10	.01	.20*	.00	-.01	.12	–		
TS	.27***	.10	.12	.15†	.23*	.07	.09	.21*	.08	.17*	.16*	.38***	.35***	.28***	.36***	.16*	.02	.22***	.01	–	
SR	.68***	.29***	.34***	.35***	.41***	.20*	.61***	.48***	.40***	.31***	.72***	.65***	.73***	.45***	.71***	.29***	.55***	.24***	.25*	.40***	–

Note. “BD” = Bear/Dragon; “BRIEF” = Behavior Rating Inventory of Executive Function; “CBQ” = Children’s Behavior Questionnaire; “:M” = Mother-report; “:F” = Father-report; “:S” = Secondary caregiver-report; “DN” = Day/Night; “DoG” = Delay of gratification (self-imposed waiting task); “FS” = Fish/Sharks; “GD” = Gift Delay; “GS” = Grass/Snow; “HG” = Hand Game; “KT” = Knock/Tap; “LiM” = Less is More; “PT” = Peg Tapping; “SH” = Shape Stroop; “SI” = Simon Says; “SD” = Snack Delay; “SST” = stop-signal task; “TS” = Token Sort; “SR” = developmentally scaled self-regulation score

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

Supplementary Table S9

ID-Wave-Rater Partial Correlations Controlling for Age

	sex	SES	SR	EXT	BRIEF	CBQ
sex	–					
SES	-.05	–				
SR	.25***	.26***	–			
EXT	-.10*	-.13*	-.10*	–		
BRIEF	.11*	.13*	.30***	-.56***	–	
CBQ	.24***	.13*	.32***	-.36***	.46***	–

Note. “SES” = socioeconomic status; “SR” = developmentally scaled self-regulation score; “EXT” = externalizing problems; “BRIEF” = Behavior Rating Inventory of Executive Function; “CBQ” = Children’s Behavior Questionnaire

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

Supplementary Table S10

EFA Factor Loadings

	One-Factor Model Standardized Factor Loading	Two-Factor Model Standardized Factor Loading	
Measure	Factor 1	Factor 1	Factor 2
Bear/Dragon	.82	.81	-.18
BRIEF: Mother	.16	.19	.44
BRIEF: Father	.31	.35	.55
CBQ: Mother	.36	.40	.55
CBQ: Father	.50	.55	.62
CBQ: Secondary Caregiver	.17	.21	.50
Day/Night	.72	.72	-.12
Fish Sharks	.63	.64	-.07
Gift Delay	.45	.44	-.14
Grass/Snow	.82	.81	-.10
Hand Game	.75	.74	-.10
Knock Tap	.89	.88	-.08
Less is More	.53	.53	.00
Peg Tapping	.83	.83	-.14
Self-Imposed Waiting Task	.64	.64	-.07
Shape Stroop	.48	.48	.12
Simon Says	.65	.65	-.20
Snack Delay	.45	.46	.11
Stop-Signal Task	.34	.35	.10
Token Sort	.51	.51	.00

Note. “BRIEF” = Behavior Rating Inventory of Executive Function; “CBQ” = Children’s Behavior Questionnaire"

Supplementary Table S11

Results from Bayesian Longitudinal Item Response Model

Predicting eta	Estimate	*SD*	Lower	Upper
intercept	1.86	4.29	-4.021	11.321
age	1.84	6.21	-7.942	15.571
age (quadratic)	11.95	6.64	3.272	28.211
female	1.81	2.04	0.004	6.330
age × female	4.01	4.34	-2.456	14.173
age (quadratic) × female	-0.05	1.95	-3.930	3.769
Predicting log(alpha)				
intercept	-2.35	0.59	-3.667	-1.309
age	-0.89	0.28	-1.311	-0.165
age (quadratic)	0.06	0.04	-0.039	0.121

Note. Significant terms are bolded. “Lower” and “Upper” are the lower and upper limits of the 95% credible interval. A coefficient would be considered statistically significant (i.e., reliably different from zero) if its 95% credible interval does not include zero.

Supplementary Table S12

Results from Model Predicting Externalizing Problems

Predictor	*B*	β	*SE*	*p*
intercept	0.20	0.00	0.02	< .001
self-regulation	-0.07	-0.28	0.01	< .001

Note. Significant terms are bolded.

Supplementary Table S13

Results from Model Predicting Externalizing Problems with Age as a Covariate

Predictor	*B*	β	*SE*	*p*
intercept	0.31	0.00	0.04	< .001
self-regulation	-0.03	-0.13	0.02	.086
age	-0.03	-0.24	0.01	< .001

Note. Significant terms are bolded.

Supplementary Table S14

Results from Model Predicting School Readiness

Outcome: Reading Skills

Predictor	*B*	β	*SE*	*p*
intercept	2.79	0.02	0.82	< .001
self-regulation	12.68	0.60	1.81	< .001

Outcome: Math Skills

Predictor	*B*	β	*SE*	*p*
intercept	6.24	0.03	0.41	< .001
self-regulation	8.69	0.79	0.47	< .001

Note. Significant terms are bolded.

Supplementary Table S15

Results from Model Predicting School Readiness with Age, Grade, and Socioeconomic Status as Covariates

Outcome: Reading Skills

Predictor	*B*	β	*SE*	*p*
intercept	-10.84	0.15	5.50	.051
self-regulation	5.63	0.27	2.21	.012
age	2.31	0.22	1.04	.029
grade	7.18	0.45	1.82	<.001
SES	0.13	0.11	0.06	.026

Outcome: Math Skills

Predictor	*B*	β	*SE*	*p*
intercept	-3.22	0.13	2.27	.157
self-regulation	5.68	0.51	0.65	< .001
age	1.85	0.34	0.47	< .001
grade	1.09	0.13	0.72	.130
SES	0.06	0.10	0.03	.022

Note. Significant terms are bolded.

Supplementary Table S16

Results from Model Predicting School Readiness with Age, Grade, Socioeconomic Status, and Intelligence as Covariates

Outcome: Reading Skills

Predictor	B	β	*SE*	*p*
intercept	-23.76	0.15	11.84	.047
self-regulation	4.05	0.19	2.30	.080
age	3.25	0.31	1.44	.026
grade	6.44	.40	1.90	< .001
SES	0.11	0.09	0.08	.153
intelligence	0.10	0.12	0.07	.172

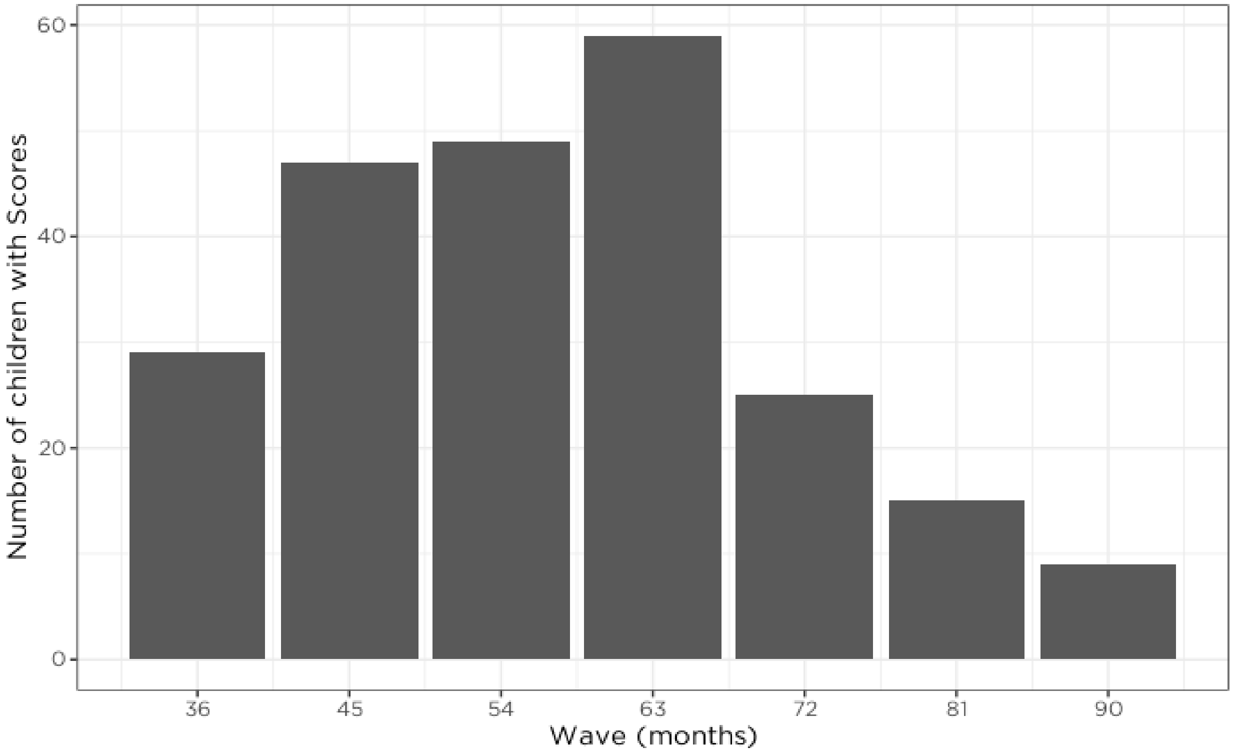
Outcome: Math Skills

Predictor	B	β	*SE*	*p*
intercept	-13.24	0.20	5.67	.021
self-regulation	3.88	0.35	1.01	< .001
age	2.47	0.45	0.62	< .001
grade	1.03	0.12	0.64	.111
SES	0.05	0.08	0.03	.104
intelligence	0.09	0.21	0.04	.036

Note. Significant terms are bolded.

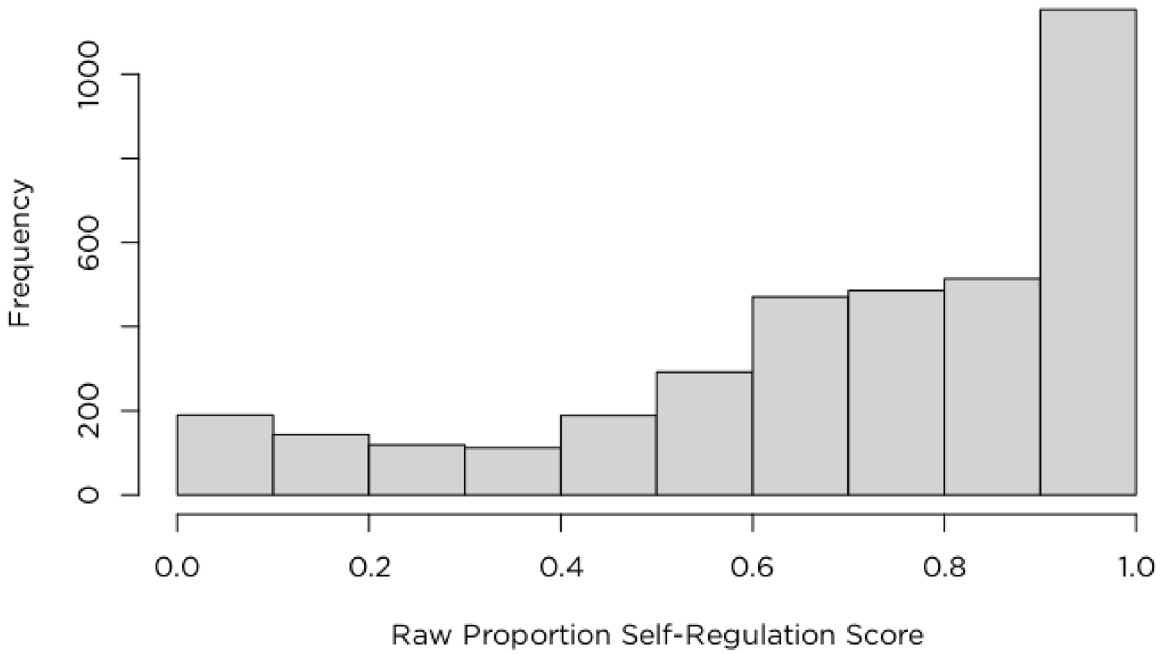
Supplementary Figure S1

Number of Children with Self-Regulation Scores by Wave (in Months)



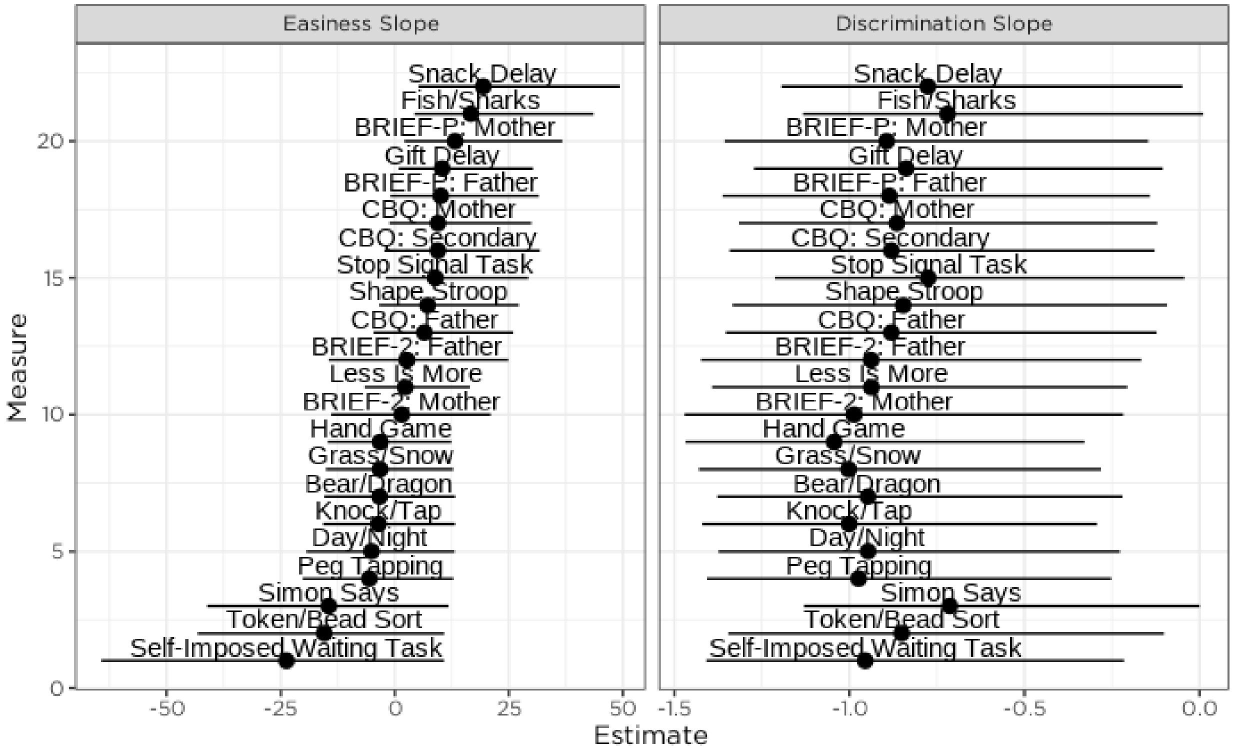
Supplementary Figure S2

Histogram of Raw Proportion Scores on Self-Regulation Measures



Supplementary Figure S3

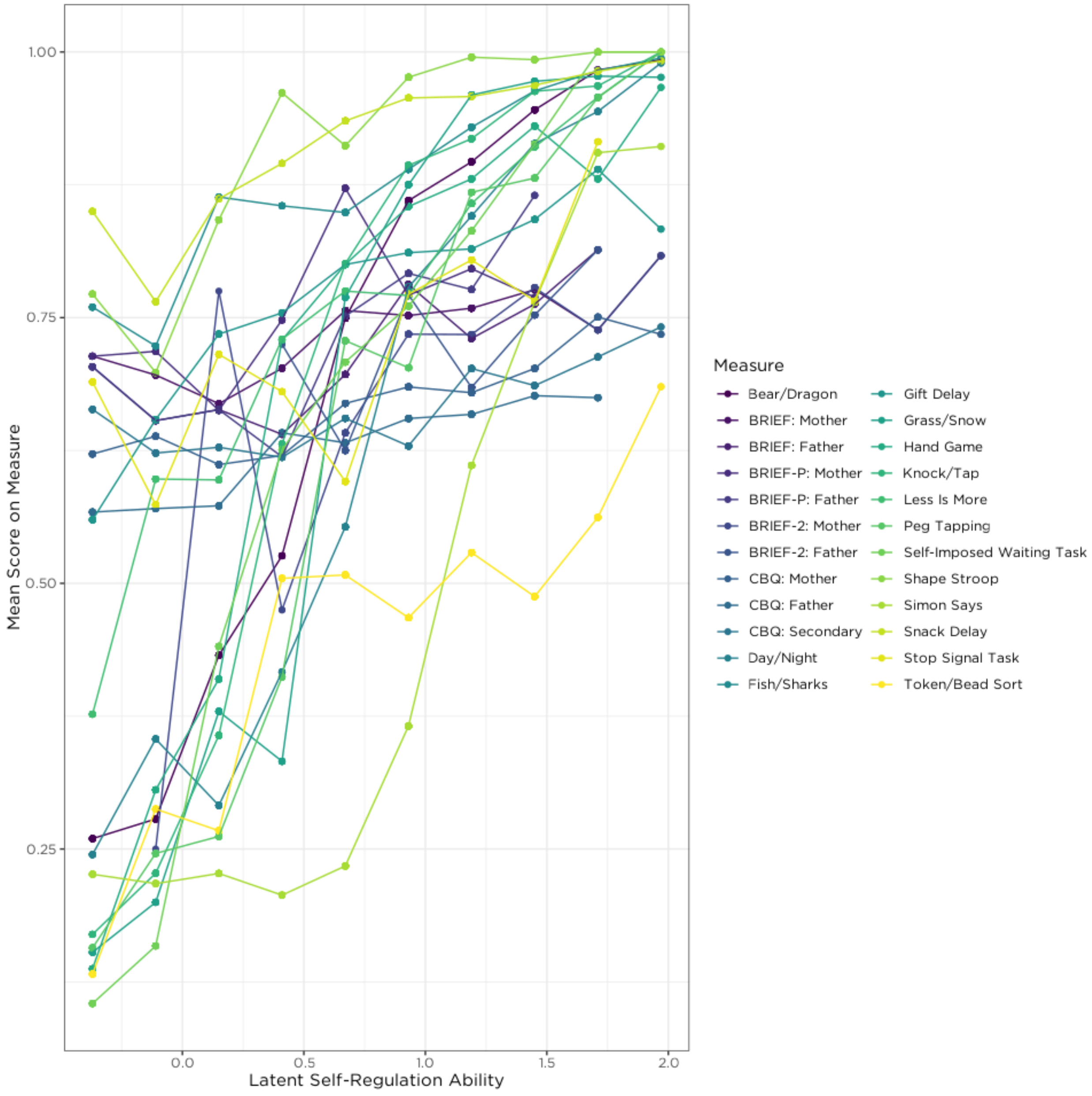
Slopes of Measures' Easiness and Discrimination Parameters as a Test of Differential Item (Measure) Functioning by Age



Note. The lines represent the 95% credible interval. "BRIEF" = Behavior Rating Inventory of Executive Function; "CBQ" = Children's Behavior Questionnaire; "Secondary" = secondary caregiver.

Supplementary Figure S4

Items' (Measures') Empirical Item Characteristic Curves



Note. "BRIEF" = Behavior Rating Inventory of Executive Function; "CBQ" = Children's Behavior Questionnaire; "Secondary" = secondary caregiver. The figure demonstrates that the measures' scores generally increased as the person's level on the latent self-regulation construct increased. Second, many of the measures showed strong discrimination (i.e., a strong association with the latent self-regulation construct, as evidenced by steep slopes). Third, the measures differed in their difficulty (location on the x-axis at the line’s inflection point). In sum, evidence suggests that the measures assessed the latent self-regulation construct well in a non-redundant way.