ELEMENTARY SCHOOLERS' DEVELOPMENT OF SELF-REGULATION: LONGITUDINAL MODELING OF EFFORTFUL CONTROL AND EXECUTIVE FUNCTION

Emily M. Weiss

A DISSERTATION

in

Education

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2023

Supervisor of Dissertation

_____

Paul A. McDermott, Professor of Education

Graduate Group Chairperson

_____

Susan A. Yoon, Professor of Education

Dissertation Committee

Michael J. Rovine, Senior Fellow of Education

Sharon Wolf, Associate Professor of Education

ACKNOWLEDGEMENT

This document, and the whole saga that is graduate school, could not have been completed without the tremendous support of some key individuals.

I would like to thank my committee members for their mentorship over the last 4.5 years and for guiding me through the dissertation process. I would especially like to thank Paul for implicitly knowing when to give me advice, when to give me space, and how to help me get out of my own way–I'm lucky to have an advisor who gets me so well. I would also like to thank Claire, Noelle and Clara for letting our writing group become a place for ongoing instrumental (and emotional!) support; it's been a pleasure to work side-by-side together towards our goals.

To my parents: thank you for helping me grow into the scholar that I am. You taught me to think like a developmentalist, even before I had the opportunity to do this work. I literally could not have done it without you.

And to Alex: you're the best teammate I could hope for. Thanks for coming with me on this journey.

ABSTRACT

ELEMENTARY SCHOOLERS' DEVELOPMENT OF SELF-REGULATION: LONGITUDINAL MODELING OF EFFORTFUL CONTROL AND EXECUTIVE FUNCTION

Emily M. Weiss

Paul McDermott

This dissertation examines children's development of self-regulation during the elementary school years by applying longitudinal measurement and latent variable models to classroom teacher ratings of children's effortful control (EC) and direct cognitive assessments of children's executive functions (EF). Self-regulation is known to support success in school, but research findings vary depending on whether capacities are defined in terms of behavior or cognition, operationalized with direct or indirect assessment, and framed in terms of developing attributes or stable traits. EC can be considered a temperamental characteristic that is theorized to manifest in relative stability over time and is often measured with adult report, whereas EF normatively improves during the early and middle childhood years and is typically assessed via performance on novel tasks. The nature of these constructs within a given child is of particular import, such that intra-individual development can be disentangled from inter-individual differences.

Therefore, this project draws on data from the national Early Childhood Longitudinal Study, Kindergarten Cohort of 2010-2011 (ECLS-K) to address three primary research objectives: (a) establish the psychometric properties of EC as assessed in the ECLS-K and generate corresponding longitudinal scores, (b) examine children's typical trajectories of change and/or stability in EC from kindergarten to fifth grade, and (c) explore heterogeneity in the co-development of EF and EC during this period.

Results show that EC has a unidimensional structure that is suitable to longitudinal linking, and can therefore be used to measure children's individual development over time. Subsequent longitudinal models indicate that EC is generally characterized by intraindividual stability rather than change, and meaningful interindividual variation in EC reveals three latent subpopulations of change that are distinguishable by level and/or shape. The final set of analyses reveal no systematic relations between growth in EF and development of EC, regardless of age group (younger versus older children) or the model employed. Findings are discussed with regard to developmental theory and methodological implications.

Table of Contents

List of Tables

List of Figures

Chapter 1 : Introduction

It is well established that the capacity to regulate one's thoughts, feelings, and behaviors in the classroom is fundamental to students' academic achievement and social wellbeing. Self-regulation (SR) is a critical component of school readiness that enables children to take advantage of classroom learning opportunities by engaging in academic instruction and developing positive relationships with teachers and peers (Blair & Diamond, 2008; Duncan et al., 2007; N. Eisenberg et al., 2010; Graziano et al., 2007; McClelland et al., 2007). Broadly defined, self-regulatory capacities meaningfully predict academic—especially mathematics—achievement (Blair et al., 2015; Duncan et al., 2007; McClelland et al., 2015; McClelland & Cameron, 2011; Zhou et al., 2010) and thus are a popular target for school-based interventions designed to improve educational outcomes and mitigate poverty-related achievement gaps (e.g., Barnett et al., 2008; Bierman et al., 2008). Nonetheless, researchers report considerable variation depending on the specific construct being measured and the nature of those measurements (Zelazo et al., 2017).

Children's capacity for self-regulation traditionally has been conceptualized in terms of effortful control (EC) or executive function (EF). EC is defined as "the efficiency of executive attention" including "the ability to inhibit a dominant response and/or to activate a subdominant response, to plan, and to detect errors" (Rothbart & Bates, 2006, p. 155). It reflects a child's behavioral tendency for attentional focusing, inhibitory control and persistence as rooted in temperament theory (Rothbart et al., 2001). By contrast, EF is a collection of neurocognitive skills that support an individual's ability to engage in goal-directed thought and action, which typically includes working memory, cognitive flexibility and inhibitory control (Zelazo et al., 2017). Whereas EC is often assessed with parent or teacher ratings of observed child behavior (i.e., indirect assessment), EF is usually measured by children's performance on novel tasks (i.e., direct assessment; Liew, 2011).

Research on childhood self-regulation is complicated by a "jingle-jangle" problem: there is considerable overlap between conceptualizations of EC and EF, yet they remain generally siloed in research traditions based respectively on behavioral psychology and cognitive neuroscience (Zhou et al., 2012). It is agreed, for example, that inhibitory control is a core component of self-regulation, yet it is operationalized very differently in EC and EF frameworks. Conflicting findings often emerge from studies employing direct or indirect assessments, leading some to suggest that they tap into different underlying constructs (Zelazo et al., 2017). However, EF performance tasks are not necessarily "pure" insofar as they can require more than one EF skill (Zelazo et al., 2017), and conceptual distinctions between EF and 'executive attention' thought to underlie behavioral manifestations of EC are not articulated in most applied work (however, see McClelland et al., 2015; Rothbart et al., 2007). Meanwhile, teacher ratings of EC derive from observations of behavior—not cognition—that can also reflect contextual influences and reporter bias (Rothbart & Bates, 2006).

Central to this debate are questions regarding stability and change. Self-regulation is generally characterized by heterotypic continuity, wherein coherent phenomena are expressed differentially across distinct phases of development (Petersen et al., 2016). Nonetheless, both personal attributes and features of the environment affect the development of cognitive and behavioral self-regulation by interacting to shape children's individual experiences (McCabe et al., 2004; see Sameroff, 2010). EF normatively improves during childhood, but distinct patterns of development may characterize different EF skills, and developmental trajectories (i.e., the rate, acceleration and timing of change) may vary considerably among children even if they attain similar levels later in life (Best & Miller, 2010; Zhou et al., 2007). Temperament theory suggests that EC is a relatively stable individual characteristic (i.e., "trait") that manifests differently over time (Rothbart et al., 2001). Nonetheless, some studies report that EC is stable (e.g., Li-Grining, 2007) whereas others find that it changes over time (e.g., Atherton et al., 2019);

the extent to which this reflects research methodology rather than substantive features of EC is not clear. In particular, teacher observations of EC may or may not appear stable across the age-graded contexts of school: because classroom expectations align with a normative level of development, relative differences over time may reflect environmental demands not intrapersonal change.

Given the importance of self-regulation in child development, experts have called to integrate EC and EF perspectives by identifying the constructs' similarities, developmental processes, and respective contributions to important outcomes (Liew, 2012; Zhou et al., 2012). Of particular interest in this paper is the nature of these constructs within a given child, such that intraindividual development can be disentangled from inter-individual differences. Such research requires data from developmentally appropriate assessments that reflect the same latent constructs over time (Petersen et al., 2021); yet, such scales are not readily available for evaluating children's EC.

1.1 Current Study

This dissertation takes an exploratory, theory-guided, measurement-focused approach to study the development of self-regulation among elementary schoolers nationwide. I draw on teacher-reports of EC (modified from the Children's Behavior Questionnaire, Rothbart et al., 2001) and direct-assessments of EF (from the Dimensional Change Card Sort, Frye et al., 1995; Zelazo et al., 2013) from the Early Childhood Longitudinal Study- Kindergarten Cohort (ECLSK) to examine children's self-regulatory development from the beginning of kindergarten to the close of 5th grade.

In order to build the foundation for future research on explanatory and causal processes, I aim to (a) establish the psychometric properties of EC as assessed in the ECLS-K and generate corresponding longitudinal scores, (b) examine children's trajectories of change and/or stability

in EC from kindergarten to fifth grade, and (c) explore the co-development of EF and EC during this period. Accordingly, I pose the following research questions (RQs) corresponding to these main objectives:

1. What is the latent factor structure of EC from kindergarten to fifth grade, as measured by teacher reports in the ECLS-K?
2. Is there evidence to support the developmental scaling of EC as measured by teacher reports from kindergarten to fifth grade under the condition of approximate measurement invariance?
3. Is EC generally stable or changing among children from kindergarten to fifth grade?
4. Is there evidence of unobserved systematic heterogeneity (i.e., multiple latent subpopulations) among children's developmental trajectories of EC during elementary school?
5. To what extent are children's developmental trajectories of EC related to those of EF, when accounting for individual differences and within-person change?
6. Is there evidence of unobserved systematic heterogeneity (i.e., multiple latent subpopulations) in the co-development of EC and EF during elementary school?

This document is organized as follows: chapter 2 reviews theory and empirical literature to provide background context for the study; chapter 3 provides an overview of methods, including details on the data source and the measures; chapters 4, 5, ad 6 cover the analysis plan and results for the three research aims noted above (RQs 1-2, 3-4, and 5-6); finally, chapter 7 includes a discussion of results, limitations, and conclusions.

Chapter 2 : Literature Review

2.1 Introduction

There is a considerable amount of empirical and theoretical work on children's self-regulation and its development. Amidst the extensive nomological net (see, e.g., Baggetta & Alexander, 2016; Bailey & Jones, 2019; Nigg, 2017), effortful control and executive function(s) are two common ways of framing enacted (behavioral, temperamental) and mental (cognitive) SR, respectively.

As introduced in the prior section, Effortful Control (EC) is thought to represent a temperamental dimension of SR that characterizes an individual's tendency for volitional and attentional (i.e., "effortful") self-control; this is in contrast to 'reactive control,' which reflects tendencies for behavioral inhibition and approach (see N. Eisenberg et al., 2005). In particular, EC is thought to leverage 'executive attention', characterized as the component of attention that is attuned to conflicting or interfering information which signals that some form of intentional control is needed in a given moment (Posner & Rothbart, 2000; Zelazo et al., 2017).

Whereas EC is conceptualized as a temperamental disposition towards self-regulatory behavior, Executive Function (EF) is generally thought to capture the cognitive foundations of self-regulation. EF is often defined as a collection of distinct but interrelated cognitive processes that integrate lower-order cognition in service to goal-directed thought or action (Baggetta & Alexander, 2016). In particular, they are often characterized as 'top-down' (Diamond, 2013), hence the name 'executive', which also corresponds to an emphasis on the "unity" (versus "diversity") of EF subcomponents (Miyake et al., 2000). Different taxonomies have been put forth, although there is a general consensus that EF–especially in adults–comprises aspects of cognitive flexibility (also called set shifting, attentional flexibility), inhibitory control (including

selective attention, interference control, response inhibition), and working memory/updating (Miyake et al., 2000); these three dimensions may not emerge distinctly among younger children, however, and the specific dimensionality of EF in early childhood under debate (Diamond, 2013; see Karr et al., 2018). Additional 'higher-order' components have also been studied, including planning, problem-solving and self-regulated learning (see Baggetta & Alexander, 2016 for review; Diamond, 2013).

There are some important areas of overlap in these two conceptualizations. Firstly, both share an emphasis on effort and top-down control. As Diamond (2013) puts it: "Using EFs is *effortful* [emphasis added]; it is easier to continue doing what you have been doing than to change, it is easier to give into temptation than to resist it, and it is easier to go on "automatic pilot" than to consider what to do next" (p. 136). In addition, Zelazo and colleagues (2017) theorize that executive attention may be a common construct underpinning both EC and EF, such that it is both "an indicator of the tendency to engage EF as well as an indicator of the temperamental predisposition to EC" (p. 53, see also Tiego et al., 2020).

Nonetheless, EC and EF traditionally have been conceptualized as separate, distinct constructs. Multiple experts have articulated how this is rooted in the history of each subfield and the corollary measures they employ; accordingly, there has been a move to integrate these subfields in order to generate more holistic conceptions of SR that are appropriate for a various developmental stages (Bailey & Jones, 2019; Diamond, 2013; e.g., Liew, 2012; Nigg, 2017; Zhou et al., 2012).

After digging into this body of literature, it is this researcher's opinion that many of the distinctions reflect not only the history of segregated research traditions, but also fundamental differences in scope and scale—that is, asking different questions and seeking answers from

different sources. Development of EF/cognitive SR is often studied in lab-based settings, focused on what tasks participants can successfully complete and fine-grained insights that can be gleaned from variations in instructions, settings, and other aspects of experimental design (see Doebel & Zelazo, 2015). In this way, there is a focus on capacity and/or ability as indicators of brain-based processes. Because much of the literature on EF and related components comes from neurocognitive research, there is also a natural interest in mapping regulatory processes to activation in distinct brain regions and networks (e.g, pre-frontal cortex, anterior cingulate). Thus, research is attuned—quite literally —to small-scale change (see Best & Miller, 2010; also Posner & Rothbart, 2000, Table 1).

By contrast, psychology literature rooted in subfields of personality, temperament, education, or applied development is generally attuned to the observable behaviors that characterize an individual over somewhat longer spans of time, especially after the rapid changes that take place during infancy. There is an inherent emphasis on individual differences, which lie at the heart of much psychological theory, and often on contextual features that vary across individuals. Put another way, the focus is not only on an individual's ability, but on the extent to which it is manifested in real-world, ecologically-valid settings. There is no doubt that activity in the brain gives rise to such behaviors, but, so far, neuroscience lacks the precision for them to be comprehensively predicted or explained. Thus different measurement traditions emerge, different research questions are prioritized, and different conclusions may be drawn.

The following review therefore focuses on literature more directly relevant to human development in the context of education. First Bailey and Jones' (2019) 'Integrative Model of Regulation' will be introduced, to provide an overarching framework with which to conceptualize SR. Although multiple theorists attempt to tease apart complexity in this topic

(e.g., Diamond, 2013; Doebel, 2020; Nigg, 2017), this particular model provides the most useful and up-to-date paradigm for conceptualizing and evaluating self-regulatory development in the context of elementary school.

Subsequently, empirical research on self-regulation and its development in the years during and adjacent to elementary school will be reviewed. This includes (i) evidence of construct validity for SR, EF, and EC, including whether EC and EF can be distinguished as separate constructs and the extent with which they are differentially associated with important school-related outcomes; (ii) what is known about stability and change in EC; and (iii) studies on the development of EF. Emphasis will be placed on literature sharing methodological overlap with the current study (i.e., work employing adult reports of children's observed behaviors, including the CBQ and TMCQ; DCCS; ECLS-K data). Each section closes with a brief summary and conclusions.

2.1.1 Theoretical Framework

Bailey and Jones' (2019) Integrative Theory of Regulation brings together research on EC and EF in a single framework, articulating the central constructs of study and when they emerge. Their integrative model posits that inhibitory control, attention control, working memory, and shifting are core regulatory processes that form the foundation of regulation in the early years of development. These core processes become enhanced with knowledge, experiences, and skills specific to cognitive, emotional, and social domains; thus skills are differentiated as lower- versus higher-order, or simple versus complex, where core processes form the foundation of more complex and/or specialized behaviors (e.g., planning, perspective-taking, prosocial behavior). Finally, the behaviors relevant to different regulatory domains

become integrated over time into a 'regulatory gestalt' that is leveraged for regulatory behaviors expected of mature individuals, like self-control and self-discipline.

They draw on principles of differentiation and integration (e.g., Overton, 2010) to articulate a general developmental sequence, which also is compatible with principle of 'stage salience'. First, generic regulatory skills emerge during toddler-hood (age 2-3), in which children begin using conscious—rather than reactive—control for the purpose of regulation. A phase of micro-level differentiation and integration characterizes the early school years (aged 3-7), where core processes become more differentiated and distinctive (reflecting the 'diversity' of EF) but also integrate to support more complex core processes; for example, shifting may represent a hierarchical integration of working memory (WM) and inhibitory control (IC), and it is generally apparent in slightly older children. Middle-to-late childhood coincides with a macrolevel phase of integration and differentiation that continues through adolescence. Between ages 7-10, core regulatory processes are fully established, and domain-specific learning allows children to leverage these core processes towards more complex ends within emotional, cognitive, or social domains; regulatory domains are generally consolidated between age 11 and 13 while domain-specific learning continues. Adolescence marks a transition towards adulthood, in which regulatory domains become increasingly integrated to being the formation of a general regulatory gestalt. Finally, the robust regulatory gestalt emerges during adulthood, such that domain specific regulatory abilities become fully integrated into a multi-faceted, transferrable capacity for self-regulation.

A few elements of their model are particularly relevant to this project. Firstly, they argue that drawing only on EF research provides a limited perspective of regulation that is "insufficient" for applied work regarding age-related changes in the first 15 years of life,

especially as it pertains to real-world situations and the contextual and social-emotional factors that influence both regulation and regulation-related outcomes (see also Doebel, 2020). The current study draws on this argument to motivate comparing developmental trajectories of both EF and EC. Secondly, the timeline they put forth suggests that children's application of regulation, as well as perhaps the subcomponents at play, undergo meaningful changes during the elementary school years; this is in keeping with the fact that later grades demand more advanced regulatory behaviors across multiple domains (e.g., group projects, longer-term due dates). Finally, the specific constructs under study in this project may be conceptualized as distinct components of their model. EF, as assessed with the DCCS, is likely a core process akin to shifting which incorporates both inhibition control and working memory (as well as the direction of attention). By nature of the assessments, EC as reflected on the teacher-reported Attention Focusing and Inhibitory Control subscales available on the ECLS-K, may in turn capture regulation in the cognitive domain, in which core processes are brought to bear in light of understanding of classroom expectations and rules.

2.2 Construct Validity of Self-Regulation

2.2.1 Dimensionality

Multiple studies have revealed a unitary factor structure for SR as measured with performance tasks and/or adult-report assessments of EC and EF among children aged 4-6 (Kälin & Roebers, 2021; Lin et al., 2019; Schmidt et al., 2022; Sulik et al., 2010). When tested, two-factor solutions distinguishing between EF and EC (or 'hot' and 'cool' measures) generally demonstrate good model fit, but fail to improve on single-factor solutions and have high interfactor correlations (Kälin & Roebers, 2021; Lin et al., 2019; Sulik et al., 2010).

Greater differentiation between SR constructs appears to emerge during middle childhood. Olson et al. (1999) interpret principal components analysis solutions as evidence of multidimensionality in impulsivity, as assessed with lab-based performance tasks; among 6- and 8-year-olds, an inhibitory control dimension emerged for both age groups. More recently, Kim-Spoon et al. (2019) found moderate a correlation (r = .27) between latent factors of EC (measured with the CBQ) and EF (measured with direct assessments) among children aged 6-8 (M age = 6.65), although a single-factor structure was not empirically tested; this value increased to r =.38 when latent factors were measured only with indicators pertaining to cognitive control (i.e., CBQ-Attentional Focusing, CBQ-Inhibitory Control, DCCS, Number Stroop task).

Similarly, in a recent systematic review and re-analysis of directly-assessed EF, Karr et al. (2018) reported that different factor structures emerged for different age groups in extant research, with greater complexity found for older children. However, their reanalysis of existing results does not reveal a conclusive structure at any age group, insofar as published models often failed to converge upon replication among bootstrapped samples. The authors suppose that this may be due to relatively underpowered studies publishing reports of more complex structures, and/or the relatively low construct reliability of the factors that emerge (i.e., weak factors with low inter-item correlations).

There is some debate about whether teacher ratings tap into the same constructs as direct assessments. For example, among a diverse sample of preschoolers in Uruguay, Nin et al. (2022) found that certain teacher-report subscales (from the Behavior Rating Inventory of Executive Function [BRIEF]) were associated with the analogous direct assessments (i.e., inhibition, planning/organizing), whereas others (i.e., working memory, flexibility) did not have significant (main effect) associations with the (hypothesized) corresponding tasks. In particular, this study

highlights that the *name* given to a subscale (e.g., "flexibility") and the specific behaviors to which it corresponds (e.g., transitioning between activities) do not necessarily hold the same meaning across subfields. It has been argued recently that tasks and surveys are fundamentally different in the extent to which they represent SR among adults, to the extent that they are not compatible with a single, over-arching latent construct; this may be due to the fact that behavioral tasks tap into a different underlying constructs than those relevant for real-world outcomes, or that they inherently lack meaningful ecological validity (I. W. Eisenberg et al., 2019; also see Enkavi et al., 2019; Friedman & Banich, 2019 for critique and response).

Relatedly, there is debate about what named constructs are tapped into with different direct assessments. Perhaps most relevant to this project are discrepancies in how to classify the Dimensional Change Card Sort, which is the focal EF measure in this study. The measure (described in detail in Chapter 3) was originally conceived to measure Theory of Mind, the ability to attribute mental states to oneself and others, which is associated with rule-following and rule-switching behavior (Frye et al., 1995). Associations between DCCS and other assessments lead to the conclusion that the DCCS must also tap into EF in some capacity; however, different authors characterize it as capturing inhibition (Best & Miller, 2010), shifting (among preschoolers, Garon et al., 2008), and/or cognitive flexibility (Zelazo et al., 2013). Doebel and Zelazo (2015) also note that the specific constructs also depend on task complexity, such that versions in which participants are asked to keep track of and follow more rules likely tap working memory and inhibitory control in addition to cognitive flexibility.

Finally, to date, only one study was identified which incorporated item-level psychometric analysis of the EC items on the ECLS-K: Husmann et al. (2022) employed an advanced longitudinal factor analytic technique (Estabrook's Measurement Model of

Derivatives) to investigate the structure of SR items, including those from the CBQ and 7 teacher-reported Approaches to Learning items. They concluded that a 4-factor structure best suited the data collected in the first 3 waves of data collection, of which one factor was primarily comprised of items originally from the Approaches to Learning scale; the remaining three factors each include 3 or 4 items from the CBQ and were termed 'Inhibitory Control', 'Distractibility' and 'Sustained Focus' by the authors (2 items did not load saliently and were excluded from confirmatory analyses). Although the authors mention that this approach can produce parameters comparable to those used in growth curve analysis (i.e., first and second derivatives terms can be transformed into intercept and slope coefficients), they did not conduct such analyses and it is not clear if resultant factors are suitable for modeling individual trajectories.

2.2.2 Predictive Validity

EC and EF also appear to differentially predict school-related outcomes in the years after children have transitioned into formal schooling. Neuenschwander et al. (2012) examined the relations between EF (performance tasks, including DCCS) and EC (parent report, CBQ) and later learning behaviors (teacher report) and academic outcomes (math, reading, and writing grades and standardized test scores) among Swiss primary-school children (*M* age 7.4 at baseline, 8.1 at follow up). They found that latent factors of EC and EF were marginally correlated ($r = .12$, $p < .10$), and both significantly predicted learning behaviors, with stronger relations observed for EF. In addition, learning behaviors were found to fully mediate the relation between EC and grades and partially mediate the relation between EF and grades, although achievement tests only were directly related to EF. Results did not meaningfully change when controlling for fluid intelligence, although substantial age effects were found for the performance-based EF measures. Their results are similar to those reported by Sánchez-Pérez et

al. (2018) who found evidence that parent-rated EC (measured with subscales from a Spanish version of the TMCQ) was associated with academic achievement only through the indirect path of Learning-Related Behaviors (organizational skills and study habits, teacher reported) among a Spanish sample of primary schoolers aged 6-12. In both cases, EC was parent-report, however, and may not perfectly reflect children's EC styles in the classroom (see section 2.3.1), recalling reports that parent- and teacher-rated EC are only moderately correlated (e.g., N. Eisenberg et al., 2009; Valiente et al., 2006)

Studies employing teacher-reports of similar construct from the ECLS-K also suggest that EF and EC in the early elementary years may differentially predict academic outcomes (e.g., Finders et al., 2021) and that learning behaviors may partially mediate the relation between EF and academic performance. In the ECLS-K, learning behaviors was assessed with 7 teacher-report items that reflect a latent construct of SR-related and motivation/engagement tendencies in the classroom ('Pays attention well', 'Works independently', 'Follows classroom rules', 'Persists in completing tasks', 'Shows eagerness to learn new things', 'Keeps belongings organized', 'Easily adapts to changes in routine'). Anthony & Ogg (2020) found that teacher-reported learning behaviors in the fall of KG partially mediated the relation between contemporaneous EF (DCCS and Woodcock-Johnson test of working memory [WJ-WM]) and growth trajectories in science through fourth grade. Indirect effect sizes were small, and the direct effects of WM and CF were larger than the indirect effects via learning behaviors and (generally) the direct effects of learning behaviors. Working memory also predicted learning behaviors (standardized $B = .24$, $p < .001$), when accounting for the simultaneous effects of the DCCS (standardized $B = .09$, $p < .05$). Similar results were found by Sung & Wickrama (2018) regarding the relation between

(linear) latent growth trajectories of EF (latent factor of DCCS and WM), learning behaviors, math and reading across kindergarten and first grade.

2.2.3 Summary & Conclusions

Overall, these findings lend empirical support to Bailey and Jones' (2019) integrative model. SR skills appear to be increasingly differentiated over time, and components are overlapping but not generally unitary. In addition, EC (and other constructs assessed via adult report) tend to reflect components beyond only EF, such as motivational, affective, and/or engagement-related processes. Furthermore, these studies show that greater EF does not necessarily result in better academic performance. Through the lens of the integrative model, this may relate to the increasing differentiation and complexity of SR tasks; it may be that improving the lower-level subskills only supports learning to the extent that they a) produce overall improvements in regulation, and b) they are not hindered by motivational or affective counterforces.

2.3 Stability and Change in Effortful Control

Empirical studies generally find evidence of rank-order stability in EC and related constructs in the years during and adjacent to elementary school. Modest individual growth is also reported, although conclusions vary more greatly across age-groups and measurement sources.

The preschool years are marked by interindividual consistency and individual growth in EC. Using lab-setting performance tasks, Kochanska & Knaack (2003) found evidence that EC, conceptualized akin to inhibitory control, had high levels of rank-order stability among very young children (r = .80 between 33 and 45 months). Similar results, albeit smaller in magnitude, were reported by Li-Grining (2007) in a study of low-income children assessed at age 2-to-4 and

again 16 months later: the author reported rank order stability of r = .40, with similar levels observed for racial/ethnic and gender subgroups, and moderate mean level changes with greater improvements at younger ages (.88, .55, and .28 *SD* for baseline 2-, 3-, and 4-year-olds, respectively).

Studies of SR in middle childhood also report relative inter-individual stability and slight or moderate mean-level growth, although the magnitude of findings depends on the construct being assessed and the type of measure that was employed. For example, significant rank-order stability in inhibitory control composites and age-related (i.e., RM-ANOVA) improvements in delay-of-gratification, behavioral inhibition, and off-task/out-of-seat behavior were reported for 6- and 8-year-olds assessed with lab-based tasks (Olson et al., 1999). High rank-order stability across 3rd, 4th, and 5th grades (r = .69-.80) was also found among parent- and child-reported EC, and growth curves revealed linear increases with significant variability in initial level and slope; no differences in growth parameters were reported for multiple-group models by age at baseline, although no evidence was provided regarding the longitudinal invariance (Lengua, 2006).

Similar results were reported for a study across three consecutive years among children in late elementary school (aged 8-12; *M* age at study entry = 9.5). Controlling for child age (i.e., age differences at baseline), King et al. (2013) found evidence of significant average increases in parent- and child-reported attention regulation and inhibitory control over time, as well as considerable variability in intercepts and slopes across participants; slope and intercept terms did not covary significantly. Nonetheless, based on descriptive statistics, it appears that the average slope was very small and that inter-individual variability was much greater than (average) intraindividual growth.

2.3.1 Reporter & Context Effects

Importantly, however, there is some evidence that the degree of stability and/or growth depends on the context in which children are observed and the role of the adult making such observations. Parent- and teacher-reports tend to be similar but not perfectly correlated, and these associations may be dependent on age.

For example, N. Eisenberg et al. (2009) found moderate correlations between parent and teacher reports of attention, impulsivity, and inhibitory control derived in part from CBQ, and smaller associations were observed for older children (r = .43 -.52 at wave 1, M age ~6 years; r = .36-46 at wave 3, M age ~9.74 yrs). Valiente et al. (2006) similarly found that parent- and teacher-reports of EC were moderately correlated among elementary- and middle-school aged children (aged 6-10 at baseline; r = .36 -.40); these associations were somewhat higher than those between adult-reports and EC observed in a laboratory setting (persistence on a difficult task; r = .09 - .16). The authors also present evidence of rank-order stability of EC over 2 and 4 years, with mother-ratings showing the greatest stability (r = .72-.83), followed by teacher-ratings (r = .49-57), and directly observed EC across 2 years (r = .54). In addition, a structural model evaluating lagged relations among children's EC, children's behavior problems, and maternal expressivity indicated a latent factor of EC that combined measurement sources was highly stable (standardized B =.75-.80), controlling for other variables in the model. However, assessor-observed EC (persistence on a difficult task) was only significantly correlated with concurrent teacher reports at the first measurement occasion, and had no significant lagged correlations with the report measures.

Similar findings were reported by Murphy, Eisenberg, and colleagues (1999) in a small sample (N < 100), 4-wave study of interindividual consistency (i.e., rank-order stability) and age-

related (i.e., mean-level) changes. Regulation was assessed at 2-year intervals beginning in preschool or kindergarten (ages 4-6) and ending in fourth, fifth or sixth grade (ages 10-12) by parent and teacher reports, primarily from the CBQ (attention focusing, attention shifting, impulsivity, and inhibition control subscales) and supplemented with a parent reports on a separate self-control rating scale; attention focusing and shifting measures changed after the earliest wave and were reported on by teachers from the second occasion onwards. The authors found high levels of rank-order stability in parent ratings during elementary school (r = .61-.81 across all regulatory outcomes), although attention focusing and attention shifting were less stable across the first and second measurement occasions (r = .25-.37) perhaps because of differences in assessments.

Teacher ratings generally revealed a lesser degree of stability: attention focusing and shifting (but not inhibitory control or impulsivity) were significantly correlated between the second and third measurement occasions (r = .34-.45), and all four teacher-reported indicators were significantly correlated between the final two waves (r = .32-60). Repeated-measures ANOVAs for the final three waves revealed that children generally showed moderate improvement in parent-observed attention shifting and inhibitory control, with girls also improving in impulsivity (boys did not change). Teachers generally did not observe growth in regulatory attributes among their elementary-school students; there was a significant (quadratic) trend only in boys' attention shifting, which increased until ages 8-10 before flattening out.

Interestingly, Zhou et al. (2007) found that teacher- and parent-reports may result in different assessments of trajectory heterogeneity among elementary-aged children. They employed the Attention Focusing subscale from the CBQ, as reported by teachers and parents, and an observed persistence assessment to evaluate the EC of children aged 5-10 on 3 occasions.

Using group-based trajectory modeling, they found evidence of substantial intra-individual stability in Attention Focusing: the three clusters that emerged differed in level but not shape (i.e., high-stable, medium-stable, low-stable). Although similar trajectory patterns emerged with data from both reporters, the classification proportions of children per class were somewhat different, such that parents tended to rate children as having lower AF than teachers did. Assessor-observed persistence revealed 3 trajectory clusters with less intra-individual stability: a high-stable class, a class that started low and rose to high by age 10, and a class with moderate baseline persistence that declined slightly over time.

Altogether these studies indicate that teachers' reports of children's EC do not necessarily provide consistent evidence of mean cohort-level improvement or intraindividual growth. For example, with regard to teacher-reported data presented by Murphy et al. (1999), the fact that mean-level change was not generally apparent and lagged correlations were only moderate could suggest that variations in students' regulatory tendencies reflect influences of the classroom setting or that differences in the timing at which classroom regulatory skills emerge generally balance out amongst the sample participants. This contrasts somewhat with reports by Zhou et al. (2007), who found that teacher-reports of attention focusing revealed stability at inter- and intraindividual levels; however, children in their study were only assessed at 3 time points, which may obfuscate nuance among individuals' trajectories as they represent age-related differences.

2.3.2 Studies of Self-Control

Somewhat more nuanced conclusions emerge from studies of self-control, which is theorized to represent a more complex aspect of SR capacities[1](Bailey & Jones, 2019). Vazsonyi & Huang (2010) found evidence of rank-order stability (standardized B = .63, .83) in mother-reported self-control between ages 4.5, 8.5 and 10.5 in a large, nationally-representative dataset. They also report significant but very gradual (mean slope = .002) increase in latent growth curves, although non-significant variability in the slope parameter suggests that everyone in their sample changed at the same rate. Nonetheless, with only three measurement occasions they could only evaluate linear trajectories, potentially masking more complex patterns of development. In this vein, Van den Akker et al. (2014) examined trajectories of personality development among Flemish youth (aged 6-20) across 5 occasions and found that mother- and child-reported conscientiousness (of which self-control is a component) was generally static during the elementary school years, with marginal average declines occurring around age 10 and a curvilinear pattern emerging during adolescence and early adulthood.

It may be that some of these complexities relate to population heterogeneity. For example, three trajectories emerged from a latent class growth analysis of teacher-rated self-control (using the Social Skills Rating System [SSRS]) from kindergarten to second grade, where half of students had above-average self-control with slightly negative quadratic curvature, a third of students had near-average self-control with a modest linear increase, and the final 15% of students had lower self-control with modest linear increases over time (Pan & Zhu, 2018). This presents a somewhat different picture than results reported by Lamont & Van Horn (2013)

[1] Although, in criminology literature 'self-control' can sometimes refer to proclivity for criminal/anti-social behavior, which considerably muddies definitions of the construct.

among a sample of low-income kindergarteners followed through 3rd grade. Using growth mixture modeling, three latent classes of parent-reported self-control (SSRS) were found, with most children (83%) having trajectories near the mean that slightly (linearly) increased over time, 10% of children initially slightly above the mean with quadratic increases reflecting greater changes occurring between 3rd and 4th and 4th and 5th time points, and only 7% of children initially slightly below the mean and having a slight increase in first three years, moderate decline after year 3, dramatic decline after year 4.

Recently, Pener-Tessler and colleagues (2022) examined heterogeneous trajectories of self-control among a large sample of twins assessed ages 3, 5, 6.5, 8-9, and 11. Four parent-reported items were included ('Keeps trying, even when something is hard'; 'Generally well behaved, usually does what adults request'; 'Thinks things out before acting'; 'Sees tasks through to the end'), and CFA and invariance testing was undertaken to validate their use in this sample. They found evidence of moderate-to-high rank-order stability (.65-.85 at adjacent occasions, controlling for measurement error), with more inter-individual variation occurring among younger children. Latent class growth analyses revealed four trajectories of quadratic change: (1) a small 'low' class did not exhibit significant change (i.e., significant intercept, nonsignificant linear or quadratic slope); (2) the largest (45%) class had high self-control without much meaningful change, although significant quadratic curvature suggests slight downward trend; (3) a small 'declining' class (20%) had slightly above average self-control at age 3 that sharply declined until age 8-9 before leveling out; and (4) an 'increasing' class (26%) had near-average levels of self-control which increased until age 8-9 before slightly decreasing. These results suggest that parent-rated self-control was relatively unchanging for just over half of the

sample, whereas the other 46% had notable periods of growth or decline during early and middle childhood.

2.3.3 Summary & Conclusions

There are a few methodological limitations that call into question the overall conclusion that EC and related constructs generally show rank-order stability (i.e., inter-individual consistency) and moderate but significant cohort-level (or intraindividual) increases over time. None of the studies described above that rely on EC questionnaire measures report tests of longitudinal invariance, therefore it is not clear whether significant growth parameters reflect average underlying changes in the latent construct or differences in the construct being measured over time.

Although some studies of self-control reported evidence of construct validity over time (Lamont & Van Horn, 2013; Pener-Tessler et al., 2022), the measures they employ do not directly inform the nature of school-based EC as framed in the current study. Furthermore, many of the studies reviewed only had access to three waves of data collection, and were therefore limited to examining only linear change (King et al., 2013; Lengua, 2006; Valiente et al., 2006; Vazsonyi & Huang, 2010). Indeed studies employing more than three time points find more complex change patterns (e.g., Murphy et al., 1999; Van den Akker et al., 2014).

An exception to these critiques is a paper by Atherton et al. (2020), who find evidence of partial strong invariance among self-and adult-reported EC questionnaires across 5 measurement occasions; nonetheless, their study pertains to youth aged 10 to 19 and thus only indirectly informs the present study. The authors found that EC (inhibitory control, attention control, activation control and composite factor) had generally high rank-order stability (r = .65 - .89 at adjacent waves). Latent growth models reveal varied patterns of change among the different

facets of EC: a linear increase in inhibitory control; a linear decrease in attention control; a nonlinear pattern in activation control with decreases from age 10 to 14, minimal change at age 16 and a substantial increase at age 19; and a similarly shaped but less extreme pattern for the composite EC factor whereby EC had only modest mean-level changes from late childhood to adolescence and increased substantially by early adulthood.

These results suggest that more nuanced conclusions about the nature of EC development can be found by employing developmentally-sensitive assessments and a large span of measurement occasions. However, until this work is undertaken for an elementary-school sample, as in the present study, it is unclear whether the EC-related behaviors that teachers observe in the classroom show rank-order consistency or variation and normative change or stagnation over time.

Finally, limited results of longitudinal mixture models suggest that there is likely considerable heterogeneity in the patterns of change exhibited by elementary school-aged children. There are conflicting findings regarding the number of latent classes (3 versus 4), whether classes vary primarily in level (e.g., Pan & Zhu, 2018; Zhou et al., 2007) or in level and shape (Lamont & Van Horn, 2013; Pener-Tessler et al., 2022), and how these trajectories play out in the school (versus home) context (e.g., Zhou et al., 2007). It is likely that contrasting results can be attributed to differences in what constructs are being assessed and how such constructs are measured, in addition to features of the sample and idiosyncrasies of modeling. Together, these leave many open questions which this project seeks to answer.

2.4 Development of EF

Whereas the degree of trait-like stability versus change in EC remains a topic of study, researchers generally agree that EF grows and changes substantially across the lifespan. A

number of factors have been suggested as drivers of EF change. Some have focused on context-independent factors, like normative neurological maturation (e.g., growth in prefrontal cortex) and genetics (e.g., Friedman et al., 2008). Others have highlighted the role of EF-relevant interactions and experiences, which is generally in keeping with bioecological, transactional, and/or developmental-systems theories of development. For example, Doebel's (2020) recent 're-thinking' suggests a reframing of EF development as comprising the emergence of control-related skills used to attain specific goals, which themselves activate foundational knowledge, beliefs, norms, values, and preferences imbued with context- and culturally-specific meaning; she omits the language of transactional developmental systems, but her arguments nonetheless align with how this framework is employed in other areas of SR research (Bailey & Jones, 2019; e.g., Olson & Lunkenheimer, 2009). Language usage (e.g., bilingualism, Park et al., 2018; private speech, Alarcón-Rubio et al., 2014), parenting practices (Sosic-Vasic et al., 2017), educational experiences (e.g., full-day kindergarten, Gottfried & Little, 2018), poverty/ SES (Blair & Raver, 2016; Haft & Hoeft, 2017; Raver et al., 2013), and race-based marginalization (Assari, 2020) have also been identified as factors that may contribute to individual differences (e.g., Zelazo et al., 2017).

Many studies employ EF as part of an explanatory framework, either looking at EF as a predictor of important outcomes, or as an outcome related to individual characteristics and/or experiences; however, a thorough *description* of EF development during the elementary school years has yet to emerge. Most of the empirical work on the development of EF employs cross-sectional models, presenting comparisons between age-groups/cohorts and differential dimensionality of EF components as evidence that EF changes in both level and structure over time (e.g., Best et al., 2011; Brocki & Bohlin, 2004). Less longitudinal work has been conducted

specifically on the developmental trajectories of EF during middle-childhood, in part because it is difficult to pin down and measure the relevant constructs; heterotypic continuity (i.e., differential manifestation of common constructs over time) creates practical hurdles for longitudinal measurement of EF (Petersen et al., 2016), and the wide range of definitions and assessments that researchers have employed has hindered attempts to draw generalizable conclusions about the patterns of EF development among school-aged children (see Baggetta & Alexander, 2016).

The following sections, therefore, review what is known about the development of EF during the elementary-school years. Because of the varied conclusions drawn with different assessments, this review focuses primarily on studies employing the Dimensional Change Card Sort (DCCS), as was used in the current project. Cross-sectional comparisons suggest that DCCS scores (incorporating accuracy and reaction time) increase from age 3 through age 15, with a cubic trajectory indicative of greater changes between ages 3 and 8 and more gradual changes until age 15, except for a slight uptick between age 10 and 12; this age-related pattern was found to explain almost 75/% of the variation in DCCS scores (Zelazo et al., 2013). There are a few longitudinal studies that shed light on the nature of EF change as measured with the DCCS. Most pertain to the first two years of school (i.e., kindergarten and 1st grade), and a few studies employed the ECLS-K data (as in this project) to examine at how EF—as measured with the DCCS and the Woodcock Johnson III Working Memory task (WJ-WM)—developed and predicted development in other outcomes. Longitudinal studies on EF development are described below, beginning with early grades (KG-1st) followed by what is known about children in grades 3-5. The section concludes with a summary of how extant findings relate to the current project.

2.4.0.1 Early Elementary School: ECLS-K Data

Willoughby and colleagues (2019) examined longitudinal association between EF and academic achievement among a full sample and children eligible for free- or reduced-price lunch (FRLP) from KG to spring of first grade. Lagged correlations between adjacent waves suggest that inter-individual stability was only moderate (r = .27-.34) using univariate and bivariate latent growth curves with structured residuals, and intra-class correlation of .26 shows that substantial variation occurs between children even though there is also a lot of variation due to individual change. Unstructured univariate growth models (i.e., latent basis) reveal that children have relatively high initial average scores (M_{G0} = 14.35, SE = .09) with significant inter-individual variability (Var_{G0}= 3.07, SE = .42), and grow on average almost 2 points by the end of 1st grade with larger increases taking place during the kindergarten year (unstandardized estimates: M_{G1} = 1.95, SE = .86, Var_{G1} = 1.95, SE = .49; Willoughby, personal communication). Secondary models added structure residuals to explain for within-person carryover in model misfit (see section 6.1 for further details). This affected the random effects, such that the intercept variance shrank from 3.07 to 1.33 and the otherwise significant slope variance (1.94) was fixed to 0; auto-regressions between the residuals were small and fixed across occasions (B = .09) in the univariate model (Willoughby, personal communication). These within-person estimates were slightly smaller than the models incorporating reading and math outcomes (B = .09-.19, depending on sample).

Concordant results have been found when examining different outcomes. Patwardhan et al. (2021) similarly used bivariate growth models with structured residuals to examine changes in DCCS and behavioral adjustment (internalizing/externalizing). Despite using a linear trajectory for growth in DCCS, their results are similar to those presented by Willoughby et al. (2019): high

average initial levels, with a linear increase of close to 2 points over 2 years, significant within-person variability in intercept but not slope, and small but significant within-person carryover effects (B = .08). These estimates were also similar to those reported by Sung & Wickrama (2018), whose model incorporated a latent factor of EF comprised of DCCS and WM scores and evaluated the simultaneous growth trajectories of EF, Approaches to Learning, and Math or Reading.

Kim et al. (2021) took a slightly different approach and examined multivariate change in DCCS, WM, and Science achievement using a random-intercepts cross-lagged panel model, controlling for demographic covariates. The between-factor model suggests moderate trait-like stability in EF between KG and spring of 1st grade (3 occasions; B = .44-.51), and significant person-level random intercepts suggest stable inter-individual differences. In addition, there was significant within-person carryover effects for both EF subskills, which were lower in value than the trait-like levels (standardized B = .12-14 for DCCS). Finally, WM and DCCS were bidirectionally related between the fall and spring of KG, but unidirectionally related (WM to DCCS) between spring of KG and spring of grade 1; parameter values were small, suggesting that more variance can be attributed to intrapersonal development (and stability) within-subskill than development across subskills.

2.4.0.2 Early Elementary School: Other Samples

Two relevant studies also employed the DCCS to evaluate EF development among other samples of young children. More nuanced findings about patterns of change in children's DCCS accuracy scores were reported by Lessing et al. (2019), who employed a latent change score model to evaluate EF development among children at age 5.5., 6, and 8. They report evidence of both reliable average change (i.e., mean of latent change score) and individual differences in

intra-individual change (i.e., variance of latent change scores). On average, larger changes occurred between the age 6 and 8 than between 5.5 and 7, although change scores between later lags had more inter-individual variation; this difference may also reflect the larger time span between waves 2 and 3. Subjects with higher EF at baseline also generally exhibited less growth by the second measurement, and greater increases in the first lag were associated with smaller increases at the later lag. Although the authors suggest that this could be the result of a ceiling effect among 8-year-olds, it seems that it may also reflect a deceleration in EF growth, such that all kids get to more or less similar levels but the timing of growth varies across individuals.

Dumont and colleagues' (2022) study provides further indication about the processes driving change in DCCS performance in the early elementary years. They found that accuracy and reaction time on the DCCS were not significantly correlated within a given occasion among children aged 5 to 7; however, cross-lagged models reveal both that significant autocorrelations emerge for both reaction time and accuracy and that reaction time was significantly associated with greater accuracy a year later, when controlling for the effects of prior accuracy. This relation was partially mediated by working memory during the first lag (i.e., age 5 to age 6). Accuracy did not predict changes in reaction time between ages 5 and 6, although greater accuracy at age 6 was associated with a lower reaction times at age 7. Collectively they interpret these findings as supporting Zelazo's 'iterative reprocessing' theory (see Doebel & Zelazo, 2015), in which EF growth comes from an emerging ability to iteratively reprocess information more efficiently by reflecting on the situation when triggered by awareness of conflict, uncertainty, or change and subsequently slowing down to reconsider options. Such reprocessing clearly embeds working memory demands, but it also seems that the ability to inhibit a prepotent response would be related, insofar as children are able to pause and reconsider. Nonetheless,

authors were not able to measure inhibitory control longitudinally, due to ceiling effects, and they employed a standard CLPM without individual-level random intercepts, which means that results highlight an overall pattern of development but do not specifically explain individual variation.

Finally, a recent study that did not employ the DCCS can help inform the current project: Reilly et al. (2022) used 6 other direct-assessment tasks to measure EF performance at 5 occasions between preschool and kindergarten, and found evidence of a nonlinear growth trajectory among in their primarily low-income sample. Children had on average much greater EF growth during pre-K than KG, with meaningful variability among the trajectories of different children. Furthermore, although children with higher EF at baseline tended to have greater growth during the preschool year, the differences attenuated such that children with relatively less growth during preschool tended to grow more during the KG year.

2.4.0.3 Late Elementary School

There is considerably less empirical research on how DCCS scores develop among children in the latter half of elementary school. As noted above, cohort-level differences in DCCS were more moderate between age 8 and 10 than between age 3 and 8 or 10 and 11 (Zelazo et al., 2013); on a more micro scale, age measured in months was also found to be linearly related to DCCS scores among 9- and 10-year-olds, although the main of age effect was quite small (standardized $B = .04$, Assari, 2020). Only one longitudinal study on the development of EF as measured with the DCCS was identified: Harms et al. (2014) looked at executive function and adaptive behavior among a community sample of children at age 8 and again at age 12 and found small-to-moderate rank-order stability in the DCCS ($r = .28$, $p < .05$). Interestingly, a

different inhibitory control task predicted teacher-rated social competence at age 12 when controlling for age 12 verbal intelligence/fluidity, although the DCCS did not.

2.4.1 Summary & Conclusions

This review of studies using the DCCS reveals a few key points. First, it is likely that change during the elementary school years is nonlinear. This is consistent with cross-sectional work by Zelazo et al. (2013) and longitudinal work by Willoughby et al. (2019), Lessing et al. (2019), and Reilly et al. (2022). It may be that the different pattern of increase during the elementary school years is due to the differential coordination of supporting EF capacities (e.g., IC and WM), as suggested by iterative reprocessing theory and empirical work by Dumont et al. (2022).

Second, it is also likely that EF is characterized, to some degree, by both interindividual stability in rank-order level and interindividual differences in the timing and/or rate of change. All of the studies described above found significant variation in baseline EF, regardless of the specific time when this was measured, and estimates of rank-order stability and/or stable interindividual differences were moderate at best (Harms et al., 2014; Kim et al., 2021; Willoughby et al., 2019). Significant variation in latent trajectory slopes was suggested by some (e.g., Sung & Wickrama, 2018); although Willoughby (personal communication) and Patwardhan et al. (2021) both fixed the slope variance to zero for their LGC-SR models, this may have been necessary for model identification, and autoregressive carryover effects reveal systematic patterns in the variation of intraindividual change processes not otherwise captured by the average trajectory. Likewise, Lessing et al. (2019) and Reilly et al. (2022) both found evidence of inter-individual differences in the timing at which more rapid growth occurred, such that children with relatively lower EF skills at one occasion mostly by exhibiting greater

increases during later lags. What remains unclear is whether all children follow a generally similar growth trajectory with slight individual variation in degree, or whether individuals' growth patterns are so varied as to reflect separable, heterogenous sub-populations characterized by qualitatively distinct archetypal trajectories.

Chapter 3 : Overview of Methods

This chapter introduces the methodology for the current study. The data and measures are described in the following sections, including descriptive statistics of key study variables. Analyses corresponding to research aims 1-3 are omitted here; specifics of the hypothesized models and analysis plans are presented alongside results in chapters 3-6.

Primary statistical analyses were conducted with R (R Core Team, 2022) and M*plus* Version 8.8 (L. K. Muthén & Muthén, 2022) using the most up-to-date versions of *lavaan* (Rosseel, 2012) and *MplusAutomation* (Hallquist & Wiley, 2018).

3.1 Data

3.1.1 Data Source

Data for this project come from the Early Childhood Longitudinal Study, Kindergarten Class of 2010-2011 (ECLS-K), published by the National Center for Education Statistics (NCES). The ECLS-K was a large scale data collection project undertaken by the US federal Department of Education, which assessed a large, nationally representative sample of 18,174 children at up to nine occasions from kindergarten to fifth grade. This project leverages the item-level response data for EC items at all waves; item-level data are crucial for establishing the factor structure of EC and creating scores appropriate for longitudinal analyses (RQs 1-2).

The ECLS-K employed a multistage stratified sampling design at baseline, whereby the country was segmented into primary sampling units (PSUs), schools were sampled within PSUs and children were sampled within schools (see Tourangeau et al., 2019). NCES reports the design effect for different weights, and the relevant value for this data (i.e., the weight accounting for attrition and sampling corresponding to child-level assessments and teacher reports; variable W9C29P_2T290) is 1.46 (Tourangeau et al., 2019, p. 4-43). Although NCES

recommends employing sampling weights to ensure results are nationally-representative, they were not incorporated into the present analyses on theoretical grounds. In effect, weights increase or decrease the extent with which each individual's data are represented in a given analysis in order to match a target population when certain groups (based on e.g., race, sex, geography) are over- or underrepresented in the sample. However, with regards to psychological constructs under study in this paper, the assumption that EF or EC will be similar among children with common demographic characteristics is not viable, and it therefore is undesirable to differentially up- or down-weight children in the sample. As an exploratory project, results are not intended to generalize as representative of the whole US population, and therefore sampling weights were not included in any stage of analyses.

The impact of clustering on this specific project could not be calculated because cluster identification variables at the classroom-level or school-level were removed from the public-use dataset to protect participant privacy. NCES also does not publish the number of participant students per classroom or school. Available information on design effects of child-level data is described below for specific variables of interest.

3.1.2 Samples

The main analytic sample for this project (N = 10,345) includes children with relevant data at 2 or more waves. Children with a parent-reported disability or teacher-reported IEP at any time point were excluded from the sample so that results reflect typically developing students. A subset of cases formed the calibration sample (N = 7000) for psychometric work corresponding to RQs 1 and 2. Random subsamples of 1000 children with complete item-level data were drawn for each wave where EC was assessed; samples were mutually-exclusive in order satisfy independence assumptions for treating each grade as a separate group (i.e., data not nested within person) as necessary for vertical linking.

3.2 Measures

3.2.1 Effortful Control

The items used in the ECLS-K to measure effortful control come from Attentional Focusing (AF) and Inhibitory Control (IC) subscales of the Children's Behavior Questionnaire (CBQ, Rothbart et al., 2001) and Temperament in Middle Childhood Questionnaire (TMCQ, Simonds & Rothbart, 2004) for younger and older children, respectively. The CBQ is a long-form measure designed to assess 15 temperament characteristics and was validated for use with children aged 3 to 7, whereas the TMCQ was developed for similar use among children aged 7 and 10 (Simonds et al., 2007). From a measurement perspective, items can only be taken to represent a theorized underlying construct if they have been validated in the exact way that they will be used (e.g., identical item wording and format, collection of items, target population, etc., Schwarz, 1999; Smith et al., 2000). Not all items from each theorized subscale were included in the ECLS-K data collection, and the language of certain items was also modified to reflect the school (versus home) context. Evidence of the dimensional structure, reliability and construct validity has been provided for the CBQ in its entirety (Rothbart et al., 2001) and as a short form (Putnam & Rothbart, 2006); analogous information for the TMCQ has not been reported in peer-reviewed journals. NCES reports alpha coefficients for the two subscales at each time point, ranging between .83 and .96 (Tourangeau et al., 2019), but the psychometric properties of the specific subset of items used in the ECLS-K have not been fully investigated.

In the ECLS-K, data on EC is available at 7 waves of data collection including the Fall of kindergarten and Spring of all grades, KG-5 (waves 1, 2, 4, 6, 7, 8, 9). Primary classroom (e.g., homeroom) teachers for grades KG-1 respond to 12 indicators of children's behavior on a 7 point scale ranging from "extremely untrue" to "extremely true" with a middle point of "neither

true nor untrue". For grades 2-5, 13 items employ a 5-point rating scale ranging from "almost always untrue" to "almost always true" with middle point "sometimes true, sometimes untrue", and are reported by children's homeroom or reading/language arts teachers, for grades 2-3 and 4-5, respectively. The documentation did not indicate which specific items correspond to each subscale (Tourangeau et al., 2019), therefore item stems were reviewed and sorted into subscales of correct length with theoretically consistent content. Table 3.2 reports the presumed subscales, the text of each item, and which items are thought to reflect common content across the CBQ and TMCQ. For interpretability, negatively worded items were reversed scaled such that higher scores indicate more self-regulation.

The design effect for certain survey components is documented by NCES. Design effects are 1.952 and 1.855 for attentional focusing and 2.400 and 2.013 for inhibitory control, at Waves 1-2 respectively. Similar information is not available for attentional focusing or inhibitory control at subsequent time points, but comparable design effects for teacher-rated Approaches to Learning are: 2.530 at Wave 1, 2.330 at Wave 2, 2.192 at Wave 3, 1.860 at Wave 4, 1.151 at Wave 5, 1.612 at Wave 6, 2.054 at Wave 7, 1.346 at Wave 8, and 1.303 at Wave 9 (Tourangeau, Nord, Lê, Sorongon, et al., 2015; Tourangeau, Nord, Lê, Wallner-Allen, et al., 2015; Tourangeau et al., 2017, 2018b, 2018a, 2019).

3.2.2 Executive Function

For the purposes of this project, EF was operationalized with a single dimension. Cognitive flexibility was assessed with the Dimensional Change Card Sort (DCCS) task from the National Institutes of Health (NIH) Toolbox for the Assessment of Neurological and Behavioral Functions (see Tourangeau et al., 2019). Originally designed to capture Theory of Mind in young children, the DCCS was extended for older participants by incorporating reaction

time. Although the ECLS-K includes additional measures of EF (working memory and visual selective attention), these data are not well suited for longitudinal analysis across the full time span in the current study. There are substantial floor effects in working memory at the first three waves of data collection (40%, 20% and 13%, respectively) that may be due in part to errors in task administration (Tourengeau et al., 2019); hence, this does not serve as a reliable source of information on young children's working memory. Visual selective attention was only measured at the final two waves of data collection and therefore cannot be examined between KG and 3rd grade (Tourengeau et al., 2019).

The DCCS includes multiple trials in which children are asked to sort images (i.e., red rabbit or blue boat) by color and/or shape, and leverages an individual's cognitive capacity to remember sorting rules and inhibit prepotent responses motivated by inapplicable sorting rules. Following practice trials, young children (< 8) began the task with 'pre-switch' trials, in which they were only asked to sort on a single dimension (e.g., color) followed by 'post-switch' trials in which they were asked to sort based on a different dimension (e.g., shape). 'Border game' trials add an additional level of complexity whereby the sorting rule (i.e., color or shape) depends on the presence or absence of a black border around the image.

For children in kindergarten and grade 1, the DCCS task was administered with physical cards and combined scores reflect the proportion of correct trials (out of 18 total). A computerized version was introduced for children in grades 2 – 5 (see Tourangeau et al., 2017). Children age ≥ 8 began with the 'mixed block' trials following practice, in which sorting rules varied by trial and certain rules were more common than others in order to induce a response tendency. Overall scores for the computerized version reflect trial accuracy (on 30 trials) and reaction time; specifically, 'overall computed' scores reflect only a child's accuracy if they

succeeded on ≤ 80% of trials but reflect the computed score derived from the accuracy rate plus reaction time if they succeeded on > 80% of trials. This scoring algorithm reflects differential 'cognitive costs' of accuracy and cognitive processing speed for younger versus older children. Scores are decomposed into 'accuracy' (range = 1.25-5, based on proportion of correct trials) and reaction time (range = 1-5) such that overall computed scores can take on values from 1.25 to 10. Changes to the testing procedure over the course of data collection arise from (i) normative developmental differences, such that older children rarely struggle with the 'pre-switch' trials, and (ii) the fact that the NIH Toolbox version of the DCCS was still being

Table 3.1

Descriptive Statistics and Correlations for Executive Function (DCCS) Scores

Occasion	*N*	*M*	*SD*	Skew	Kurtosis	Pct. Max.	1	2	3	4	5	6	7	8
1. T1	9110	4.3	0.65	-1.71	2.3	8.32	—	—	—	—	—	—	—	—
2. T2	10166	4.5	0.53	-2.06	5.3	16.77	.27	—	—	—	—	—	—	—
3. T3	3334	4.5	0.46	-2.14	6.7	24.33	.24	.26	—	—	—	—	—	—
4. T4	9421	4.6	0.43	-2.38	8.4	31.43	.21	.20	.28	—	—	—	—	—
5. T5	3058	6.5	1.34	-1.40	1.9	0.00	.25	.25	.23	.24	—	—	—	—
6. T6	8742	6.8	1.23	-1.55	2.9	0.01	.20	.19	.19	.23	.51	—	—	—
7. T7	8073	7.3	0.97	-1.44	4.5	0.01	.16	.14	.12	.18	.41	.42	—	—
8. T8	7563	7.7	0.85	-1.10	4.5	0.11	.14	.11	.14	.15	.36	.38	.43	—
9. T9	7128	8.1	0.84	-0.83	3.4	0.45	.13	.12	.13	.14	.31	.33	.39	.45

Note.
Pearson-product moment correlations calculated with pairwise complete observations. Scores at T5-T9 represent accuracy and reaction time. Scores at T1-T4 reflect total accuracy only and have been linearly transformed for consistency. Pct at Max. denotes the percentage of students with available data who obtained the maximum score (5 at T1-T4; 10 at T5-T9).

developed when ECLS-K data collection began in 2010. This project employed the overall scores noted above, since case-wise, trial-level data are not available in the public-use dataset.

Descriptive statistics for the DCCS scores used in this study are presented in Table 3.1. For comparability, scores for T1-T4 were linearly transformed to match the accuracy component

of computed scores at T5-T9 (i.e., 1.25-5). Distributions are considerably nonnormal at certain waves, including large percentages of children at T3 and T4 who succeed at all of the trials on which they were tested.

Table 3.2

Effortful Control Item Stems by Theorized Subscale

Items taken from the CBQ

Attentional Focusing

1. [a]When practicing an activity, has a hard time keeping his/her mind on it. *R* 2. Will move from one task to another without completing any of them. *R*
3. When drawing or coloring in a book, shows strong concentration.
4. When building or putting something together, becomes very involved in what s/he is doing, and works for long periods.
5. [b]Is easily distracted when listening to a story. *R*
6. Sometimes becomes absorbed in a picture book and looks at it for a long time.

Inhibitory Control

7. [e]Can wait before entering into new activities if s/he is asked to.
8. Plans for new activities or changes in routine to make sure s/he has what will be needed.9. Has trouble sitting still when s/he is told to (story time, etc.). *R*
10. [c]Is good at following instructions.
11. Approaches places that s/he thinks might be "risky" slowly and cautiously.
12. [d]Can easily stop an activity when s/he is told "no".

Items taken from the TMCQ

Attentional Focusing

1. [b]Is easily distracted when listening to a story *R*
3. Looks around the room when doing school work. *R*
5. [a]When working on an activity, has a hard time keeping her/his mind on it. *R*
7. Has a hard time paying attention. *R*
9. Needs to be told to pay attention. *R*
10. Gets distracted when trying to pay attention in class. *R*

Inhibitory Control

2. [d]Can stop him/herself when s/he is told to stop
4. Can stop him/herself from doing things too quickly
6. [e]Has an easy time waiting
8. Has a hard time waiting his/her turn to talk when excited. *R*
11. Likes to plan carefully before doing something.
12. [c]Is good at following directions.
13. Has a hard time slowing down when rules say to walk. *R*

Note:

Items sharing the same superscript are thought to have common content. *R* denotes reverse coded

CBQ = Children's Behavior Questionnaire

TMCQ = Temperament in Middle Childhood Questionnaire

Chapter 4 : Psychometric Analysis of EC

The first research aim comprises psychometric analyses of EC as measured in the ECLSK. As noted in previous chapters, utilizing questionnaires without verifying the specific constructs under study for the target population (or sample) can be a substantial threat to validity. Hence, before longitudinal modeling of self-regulation can proceed, the following research questions were addressed: (RQ1) *What is the latent factor structure of EC from kindergarten to fifth grade, as measured by teacher reports in the ECLS-K?*; and (RQ2) *Is there evidence to support the developmental scaling of EC as measured by teacher reports from kindergarten to fifth grade under the condition of approximate measurement invariance?*

4.1 Methods

4.1.1 RQ1

Psychometric analyses of EC began by examining confirmatory models reflecting the subscale structure of the CBQ and TMCQ. As previously described, items for each age-block were drawn from two of the three subscales comprising a higher-order dimension of EC. At this stage, confirmatory factor analyses were conducted separately for CBQ and TMCQ items (waves
1, 2, 4 versus 6, 7, 8, 9), and it was not presumed that the same factor structure would necessarily best capture the dimensionality of EC as measured with each set of items.

It was hypothesized that one of the following four models would characterize each ageblock: (1) a unidimensional model (Model 1.1) reflecting the single dimension of EC, (2) a unidimensional model with residual covariances accounting for subscale/wording effects (Model

1.2), (3) a correlated two-factor model reflecting the theorized subscales of AF and IC (Model 1.3), and (4) a bifactor model (Model 1.4) in which a general EC factor is distinguished from specific factors of AF and IC (see Reise et al., 2007). Models 1.1 and 1.3 reflect the assumption of local independence, whereas model 1.2 accounts for dependence among items with similar language that is not accounted for by the common factor(s). In the present study, the bifactor model (Model 1.4) accounts for specific factor common variance that is presumed to be both reliable and meaningful (i.e., not 'nuisance' factors). It represents a "direct hierarchical" relationship (Gignac, 2008) between EC and the two subcomponents insofar as the associations between individual items and EC are not mediated by latent factors of AF and IC; although the subscales are thought to represent a higher-order factor, a higher-order model with only two lower-order factors is not identified and therefore cannot be evaluated.

4.1.1.1 Analysis Plan & Evaluation Criteria

Competing models were evaluated against multiple criteria. These include: (1) structure consistent with theory, including salience of factor loadings (>.40), and latent factor covariance <.85 for Model 1.3 (Brown & Moore, 2012); (2) nested model comparisons via chi-square difference testing, with appropriate corrections for non-normality (Satorra & Bentler, 2010); and (3) approximate fit indices (AFIs) wherein SRMR < .08, comparative fit index (CFI) >.95 and root mean square error of approximation (RMSEA) <.06 are indicative of good fit, and CFI >.90 and RMSEA < .07 are indicative of adequate fit (Hu & Bentler, 1999). In light of recommendations to avoid employing fixed thresholds for ascertaining model fit (e.g., Brown & Moore, 2012; Marsh et al., 2004), AFIs were employed to gauge the relative improvement of different models, with the values above as guidelines. Chi-square test statistics often reject correctly-specified models among large samples (Brown & Moore, 2012), thus it was expected

that this would indicate a significant discrepancy between the model and the data. Resultant scale(s) were evaluated for evidence of sufficient reliability, including omega coefficients appropriate for the given model (Rodriguez et al., 2016).

Data were treated as continuous, as recommended for ordinal data with 5 or more response categories, and estimated via maximum likelihood with robust corrections to account for non-normality (Finney, S.J; DiStefano, 2006; Robitzsch, 2020c). In addition, models were re-run treating data as categorical via diagonally-weighted least squares (i.e., WLSMV), in order to ensure the robustness of results to different estimators. Appropriately scaled fit indices were examined for each estimator.

4.1.2 RQ2

In order to produce factor scores that can be used for longitudinal measurement, it is essential that latent factors be made comparable across time points. Historically it was argued that this can only be done under the condition of strict measurement invariance, wherein factor loadings, intercepts, and residual variances are equal for all groups or time points (see Millsap & Olivera-Aguilar, 2012). Others suggest that partial measurement invariance is sufficient, which involves constraining invariant parameters to equality and allowing noninvariant parameters to be freely estimated within a given group or time point (Byrne et al., 1989; Millsap & Kwok, 2004); however, this requires detecting truly invariant parameters (i.e., anchor items) to be used to identify the model (Rensvold & Cheung, 1998), which can be difficult to ascertain without prior knowledge (Kopf et al., 2015; Lai et al., 2021; M. Wang & Woods, 2017). An alternative framework through which to view this dilemma is that of approximate invariance (Asparouhov & Muthén, 2014; B. O. Muthén & Asparouhov, 2013). In contrast to exact invariance, in which model parameters are either equal across groups or freely estimated, approximate invariance

tests the weaker assumption that parameters are close enough to be considered approximately equal (Lai et al., 2021).

The alignment method, introduced by Asparouhov and Muthen (2014), produces factor loading and intercept parameters that can be compared across groups under the assumption of approximate invariance and provides a more parsimonious alternative to searching for anchor items. This method minimizes the difference between a configural model, in which values for each group are separately estimated, and an aligned model with equal model fit (i.e., likelihood value). It does this by establishing a baseline model with standard normal factor means and variances within each group and subsequently estimating loading and intercept parameters that satisfy a simplicity function (Asparouhov & Muthen, 2014, Equation 9) by which original and transformed parameter values are most similar (see Asparouhov & Muthen, 2014). The authors note that this is similar to rotation in exploratory factor analysis (EFA) insofar as it obtains a solution with many invariant parameters and few highly noninvariant parameters, rather than many parameters with moderate degrees of noninvariance. It can be accomplished with maximum likelihood or Bayesian estimation, and the authors suggest that Bayesian approaches may be more robust. Although it was originally developed as an exploratory procedure for comparing multiple groups, it has been extended to confirmatory approaches, termed alignment-within-CFA (AwC; Marsh et al., 2018), and applied to form the measurement model for second-order growth curves (Lai, 2021). Most relevant to this project, the alignment method also has been reframed as a means of scale linking in an IRT context (Pokropek et al., 2020), due to its similarities with the Haberman linking method previously developed (Haberman, 2009; see Robitzsch, 2020a for comparison).

4.1.2.1 Analysis Plan & Evaluation Criteria

The alignment method was therefore applied to obtain model parameters for a full longitudinal model of EC based on the final model ascertained in RQ1. First, configural models were estimated, employing mutually exclusive random subsamples of each grade level (from the calibration sample) to satisfy independence assumptions. To aid interpretability and comparability of the two sets of items, the CBQ items were rescaled to match those of the TMCQ (i.e., 1-5), and items deemed to have common content (noted in Table 1) were treated as applicable to both sets of items (i.e., each treated as a single item). Subsequently, the alignment optimization method was used to obtain a linked model, as described by Pokropek et al. (2020) and Robitzsch (2020b). As noted above, testing for differential item functioning is not necessary in this paradigm, as the alignment method finds the optimal solution via rotation; common content items with aligned parameters overcome the problem of factor indeterminacy that would otherwise arise.

The quality of alignment optimization for longitudinal linking was evaluated by establishing the number of invariant items and percentage of invariant parameters (i.e., loadings, intercepts) for common content items. Because not all items were administered at all waves, the common content items are most relevant for linking the latent factor across grades. Significant noninvariance was determined via the automated algorithm in Mplus, which relies on a two-step process of pairwise comparisons followed by comparisons of individual parameters against an invariant set (Asparouhov & Muthén, 2014). In addition, the magnitude of parameter differences was examined (as per Lai, 2021; Luong & Flake, 2022) to ascertain the degree of (non)invariance.

Recommendations for how many noninvariant parameters are acceptable vary based on the characteristics of the simulation studies that have been conducted. Muthen and Asparouhov (2014) suggest that fewer than 25% of parameters is acceptable for making group comparisons in latent factor means; a similar level (29%) was suggested by Flake and McCoach (2018) regarding categorical data with a two-factor solution among two groups, and these recommendations were echoed by Luong and Flake (2022) in their applied tutorial. Lai (2021) found that for a longitudinal model with 5 items across 4 time points, 25% noninvariant parameters performed acceptably whereas 55% noninvariant parameters did not. Simulation studies with a wider variety of conditions have put forth less stringent recommendations: Lai and colleagues (2021) found that 33% noninvariant parameters produced unbiased results, and even 67% noninvariant parameters was acceptable when the direction of parameter differences approximately balanced out. They conclude that their findings are therefore more similar to those by Marsh et al. (2018) who found support for a model even with 100% noninvariant parameters when the degree of noninvariance was small for all but a few parameters.

Finally, longitudinal factor scores were produced for the full analytic sample based on the parameters obtained from the aligned calibration sample model.

4.2 Results

4.2.1 RQ1

As described above, four a priori hypothesized models were tested to evaluate the dimensional structure of the EC items obtained from the CBQ and TMCQ. Analyses were conducted with data from the calibration sample separately for waves 1,2 and 4 (i,e., CBQ items) and waves 6-9 (i.e., TMCQ items). Primary analyses were conducted with *lavaan*, employing the 'MLM' estimator and 'mimic = MPLUS' specification in order to match what

would have been produced via Mplus (see Savalei & Rosseel, 2022). Model fit statistics are presented in Table 4.1 and parameter estimates are given in Tables 4.2 and 4.3.

Table 4.1

Fit Statistics of CFA Models

Model		$\chi^2(df)$	CFI	RMSEA	SRMR
CBQ					
	Mod1.1.1	2367 (55)	0.85	0.14	0.14
	Mod1.2.1	518 (38)	0.97	0.07	0.03
	Mod1.3.1	1828 (53)	0.88	0.12	0.06
	Mod1.4.1	605 (42)	0.96	0.08	0.03
TMCQ					
	Mod1.1.2	2582 (65)	0.92	0.11	0.06
	Mod1.2.2	521 (40)	0.98	0.06	0.07
	Mod1.3.2	1753 (64)	0.94	0.09	0.04
	Mod1.4.2	2299 (55)	0.93	0.11	0.20

Note:
Analyses conducted with calibration sample.
Statistics are scaled to account for non-normality.

For both the CBQ and TMCQ, Model 1.2 was selected as best-fitting and retained for further analysis. For the CBQ, results indicate that both the unidimensional model (Model 1.1) and the two-factor model (Model 1.3) reflecting the hypothesized structure of the CBQ subscales were quite ill-fitting, both with CFI less than .90 and RMSEA >.10. Analogous models for the TMCQ revealed 'adequate' if not quite 'good' fit (RMSEA >.08, CFI <.95); however, the interfactor correlation between Attentional Focusing and Inhibitory Control subscales for Model 1.3 was very high (TMCQ = 0.89). This was also the case for the CBQ (CBQ = 0.92), suggesting that the two subscales may be more reflective of a single underlying construct for both sets of data. The remaining models evaluate whether the items reflect a single unidimensional factor of

EC after accounting for variance pertaining to the respective subscales. The one-factor model with residual correlations accounting for similar wording for each subscale (Model 1.2) clearly offered the best model fit for the TMCQ items (CFI = 0.99, RMSEA =0.06) and also fit the CBQ

Table 4.1

CBQ Parameter Estimates

	Model 1.1.1	Model 1.2.1	Model 1.3.1		Model 1.4.1		
	EC	EC	AF	IC	EC	AF	IC
Inhibitory Control							
cbq1	**0.63**	**0.8**	**0.83**		**0.85**	-0.13	
cbq2	**0.72**	**0.75**	**0.79**		**0.8**	-0.1	
cbq3	**0.58**	**0.59**	**0.63**		**0.63**	**0.45**	
cbq4	**0.53**	**0.53**	**0.58**		**0.59**	**0.6**	
cbq5	**0.76**	**0.83**	**0.81**		**0.82**	-0.07	
cbq6	**0.42**	**0.42**	**0.46**		**0.45**	**0.41**	
Attentional Focusing							
cbq7	**0.62**	**0.61**		**0.68**	**0.59**		0.33
cbq8	**0.67**	**0.68**		**0.71**	**0.69**		0.13
cbq9	**0.77**	**0.83**		**0.8**	**0.77**		0.22
cbq10	**0.81**	**0.81**		**0.86**	**0.8**		0.29
cbq11	0.37	0.36		**0.42**	0.34		0.3
cbq12	**0.72**	**0.72**		**0.79**	**0.66**		**0.62**
Reliability Coefficients							
omega	0.68	0.87	0.85	0.86	0.87	0.04	0.16
alpha	0.91	0.91	0.85	0.86	0.91	0.85	0.86

Note:
EC= Effortful control. AF = Attentional Focusing. IC = Inhibitory Control. Bold entries are salient (>.40).

items quite well (CFI = 0.97, RMSEA = 0.07). Although the bifactor model (Model 1.4) offered comparable fit to the residual correlations model for the CBQ items, it exhibited relatively poorer fit for the TMCQ items. Model 1.2 likewise offered the best fit for both the CBQ and TMCQ data when employing alternative means of accounting for nonnormality (i.e., MLR and WLSMV estimation; see Table 4.9).

Likelihood ratio tests of nested models were also used to compare models. Following the nesting and equivalence testing procedure outlined by Bentler and Satorra (2010; as implemented in the semTools R package, Jorgensen et al., 2022), four nesting relations could be tested for the CBQ and one could be tested for the TMCQ. All tests were statistically significant, indicating that the less parsimonious models were worth the additional parameters (see Table 4.4).

Table 4.2

TMCQ Parameter Estimates

	Model 1.1.2	Model 1.2.2	Model 1.3.2		Model 1.4.2		
	EC	EC	AF	IC	EC	AF	IC
Inhibitory Control							
tmcq1	**0.83**	**0.82**	**0.83**		**0.64**	**0.64**	
tmcq3	**0.82**	**0.79**	**0.82**		**0.73**	**0.45**	
tmcq5	**0.87**	**0.82**	**0.88**		**0.76**	**0.5**	
tmcq7	**0.91**	**0.86**	**0.92**		**0.79**	**0.5**	
tmcq9	**0.88**	**0.85**	**0.89**		**0.79**	**0.46**	
tmcq10	**0.92**	**0.87**	**0.92**		**0.8**	**0.5**	
Attentional Focusing							
tmcq2	0.39	**0.42**		**0.45**	**0.41**		**0.8**
tmcq4	**0.44**	**0.48**		**0.52**	**0.5**		**0.41**
tmcq6	**0.6**	**0.66**		**0.68**	**0.69**		0.13

tmcq8	**0.61**	**0.65**		**0.66**	**0.68**		0.05
tmcq11	**0.68**	**0.72**		**0.72**	**0.74**		0.02
tmcq12	**0.79**	**0.83**		**0.83**	**0.84**		0.03
tmcq13	**0.63**	**0.68**		**0.69**	**0.71**		0.05
Reliability Coefficients							
omega	0.92	1	0.95	0.81	1.04	0.45	0.1
alpha	0.94	0.94	0.95	0.84	0.94	0.95	0.84

Note:
EC = Effortful control. AF = Attentional Focusing. IC = Inhibitory Control. Bold entries are salient (>.40).

Table 4.3

Likelihood Ratio Tests of Nested CFA Models

H_0	H_1	$\Delta\chi^2$	DF	p
CBQ				
1.1	1.2	1784	17	0
1.1	1.3	383	2	0
1.1	1.4	1896	13	0
1.3	1.4	1269	11	0
TMCQ				
1.1	1.3	285	1	0

Note:
Nesting relations and LRT following Bentler and Satorra (2010); MLM estimation.

Inspection of the factor loadings for Model 1.2.1 reveal that a single item was not salient for the CBQ data; therefore, a secondary model was evaluated without item 11 ("Approaches places that s/he thinks might be 'risky' slowly and cautiously"). This modified model provided comparable fit (scaled χ^2 = 482.05, CFI = 0.97, RMSEA = 0.08, SRMR = 0.03) with only marginally higher reliability coefficients (Omega = 0.88, Alpha = 0.92). Because removing the item did not appreciably change the model or latent factors, it was decided to retain the item in

subsequent models to maintain consistency with extant research. Therefore, the originally hypothesized Model 1.2.1 was employed in subsequent analyses.

4.2.2 RQ2

4.2.2.1 Preliminary Analyses

Prior to conducting the alignment optimization, the final CFA model was reproduced separately for each measurement occasion. These models all display adequate or good model fit (Table 4.5), confirming that the factor structure is suitable for the alignment method. Inspection of parameter estimates show that a few items had non-salient factor loadings as noted above: cbq11 for kgS (0.317), cbq6 for g1 (0.295), cbq11 for g1 (0.343), tmcq2 for g3 (0.359), and tmcq2 for g5 (0.372). These items were retained in order to maintain a consistent factor structure across grades and comport with extant research employing this data; because none of these reflect common content their presence does not affect the results of the alignment optimization procedure.

Table 4.4

Fit Information for Single-Grade CFAs, Model 1.2, MLR

Grade	Chi-square	DF	CFI	RMSEA	SRMR
KG fall	191	38	0.97	0.07	0.03
KG spring	208	38	0.97	0.08	0.04
1st	205	38	0.97	0.07	0.04
2nd	141	40	0.99	0.06	0.09
3rd	158	40	0.99	0.06	0.03
4th	196	40	0.98	0.07	0.10
5th	172	40	0.98	0.06	0.10

4.2.2.2 Alignment Method

The alignment optimization procedure resulted in 8 (11.43%) noninvariant parameters for the common content items, comprising the Intercept for item CB5_TMC1 at Fall of KG, Spring of KG, and 1st grade, the Intercept for item CB7_TMC6 at Spring of KG, and 1st grade, and the Loading for item CB7_TMC6 at Fall of KG, Spring of KG, and 1st grade. These results fall well into the guidelines recommended by Lai et al. (2021), Asparouhov and Muthen (2014) and Flake and colleagues (Luong & Flake, 2022; Flake & McCoach, 2018). The magnitude of parameter differences was generally small, on average 0.187 for loadings and 0.017 for intercepts (respectively 3.741 and 0.346% of the original 5-point scale).

In light of these findings, a secondary model was tested using a nondefault loss function where the simplicity was set to the fourth root, which may result in smaller parameter differences in some cases (B. O. Muthén & Muthén, 2012). This revealed a slightly different pattern of noninvariant parameters, substituting the Loading for item CB12_TMC for that of CB7_TMC6, and including the Intercept for item CB1_TMC5 at 1st grade in addition to the intercepts previously detected; parameter estimates are presented in Table 4.6. Although this second model had a one additional noninvariant parameter (i.e., 12.86% noninvariance), the magnitude of the differences was generally smaller, resulting in group differences of on average only 0.003 for loadings and 0.024 for intercepts (0.069% and 0.489% of the original 5-point scale, respectively), with a pairwise maximum of only 0.234 (4.68% of original scale). These results suggest that the alternative specification of the loss function was successful in reducing parameter differences, and therefore this secondary model was retained for further investigation.