Clinical Child and Family Psychology Review (2019) 22:63–74
https://doi.org/10.1007/s10567-019-00285-1

Measuring Young Children's Executive Function and Self-Regulation in Classrooms and Other Real-World Settings

Dana Charles McCoy[1]

Published online: 18 February 2019

Abstract
A number of different approaches are currently used for assessing young children's executive function (EF) and self-regulation (SR) skills. Nevertheless, guidance for stakeholders aiming to assess EF and SR in real-world settings (e.g., preschool classrooms) is currently lacking. In the present article, I review the properties, strengths, and weaknesses of three common approaches to EF and SR measurement: direct assessments, adult reports, and observational tools. Building on this general review, I next highlight several considerations specific to EF and SR measurement of young children in everyday contexts. In particular, I consider the ecological validity, interpretability, and scalability of each approach to EF and SR measurement, concluding with future directions for research.

Keywords Self-regulation · Executive function · Early childhood · Measurement · Ecological validity

Introduction

A large body of research has shown children's early executive function (EF) and self-regulation (SR) skills to be foundational to their later-life success. EF and SR in early childhood have been linked with a host of important outcomes, including improved academic performance and health, increased adult earnings, and reduced criminal behavior (Blair and Razza 2007; Moffitt et al. 2011; Schlam et al. 2013). Importantly, evidence also shows that EF and SR skills are malleable, and a number of early interventions designed to improve these capacities have shown lasting impacts on a range of developmental domains (Diamond and Lee 2011; Pandey et al. 2017; Raver et al. 2011).

Historically, a variety of approaches from different disciplines and settings have been used to operationalize EF and SR. These approaches reflect a diversity of goals related to understanding EF and SR in research and clinical practice. Despite the broad availability of EF and SR assessments, relatively few measures have been developed with the specific intention of capturing young children's EF and SR skills *in context*, in the everyday environments in which children must learn to pay attention, avoid impulsive reactions, regulate emotions, and plan and monitor their behavior. In particular, despite a growing interest in early education-based intervention programming, relatively little is known about how best to capture preschoolers' EF and SR skills in the context of the classroom environment.

The goal of the present paper is to inform researchers, clinicians, and educators regarding the relative advantages and disadvantages of existing approaches to measuring EF and SR during the early childhood period, including direct assessment, adult report, and observational approaches. Building on this foundational literature, I also aim to highlight several considerations specific to operationalizing children's EF and SR skills in applied settings, such as classrooms. In particular, I emphasize the importance of *ecological validity*, or a measure's ability to capture the ways that children deploy EF and SR skills in the context of real-world distractions, emotions, and supports. Furthermore, I highlight the importance of generating *interpretable* data that are directly tied to decision-making, as well as approaches that are *scalable* for large-scale use. Doing so, I argue, will provide more actionable information to researchers and practitioners interested in supporting children's EF and SR behaviors in the everyday environments in which they learn, grow, and thrive.

✉ Dana Charles McCoy
dana_mccoy@gse.harvard.edu

[1] Harvard Graduate School of Education, 704 Larsen Hall 14, Appian Way, Cambridge, MA 02138, USA

Existing Approaches to Measuring Executive Function and Self-Regulation

Executive function and self-regulation have been defined in a number of different ways, using a variety of different disciplinary and conceptual frameworks (Jones et al. 2016). For the purposes of this paper, self-regulation (SR) is defined as a broad set of both conscious and unconscious processes that individuals use to regulate (e.g., control, modulate, inhibit, initiate) both their internal states (e.g., attention, emotion) and observable behavior (McCoy 2013; Nigg 2017). Executive function (EF) can be thought of as a set of higher-order cognitive processes that facilitate "top down" SR, including set-shifting, working memory, and inhibitory control (Miyake et al. 2000). Because EF and SR are often discussed hand-in-hand in the early childhood literature, I use both terms throughout the rest of this review.

Although theorists such as James, Piaget, and Vygotsky have described regulatory phenomena for over a century (Fox and Riconscente 2008), formal quantitative measures of children's EF and SR are relatively new (Zelazo et al. 2016). In the 1970s, Walter Mischel introduced his now-famous "marshmallow task", which captures children's delay of gratification by measuring how long they are able to wait to eat a desirable snack and what strategies they use to avoid temptation (Mischel et al. 1972). Over the past several decades, at least 80 new measurement tools have emerged for representing different dimensions of children's EF and SR (Ackerman and Friedman-Krauss 2017). These approaches range on a number of characteristics, including their conceptual foci, disciplines of origin, and intended setting. In Table 1, I review the characteristics of three primary approaches to measuring EF and SR in modern early childhood research, including a brief discussion of the relative advantages and disadvantages of each.

Direct Assessments

Direct assessments are defined for the purpose of this paper as performance tasks (often described to participants as "games") that are administered directly to children either one-on-one or in a group setting to assess specific sub-constructs of EF and SR. Although many direct assessments were originally developed for laboratory-based psychological research or clinical assessment, an increasing number of these tasks have been adapted for use in the field. In general, direct assessments tend to be relatively fast to administer, taking approximately 2- to 6-min each to complete, excluding setup time. Because of their long history as research and clinical tools, the psychometric properties of many direct assessment tools have been studied extensively, providing positive evidence for their reliability, validity, and metric invariance (e.g., Willoughby et al. 2012; Zelazo et al. 2016). They have also been used widely in the context of evaluation studies, though evidence regarding their sensitivity to early intervention impacts is somewhat mixed (Morris et al. 2014; Raver et al. 2011; Weiland and Yoshikawa 2013; Zelazo et al. 2016).

The marshmallow task is a classic example of a direct assessment of effortful control—a set of EF-like inhibitory and attentional skills that involve the management of emotions (Jones et al. 2016)—and has been joined by

Table 1 Summary of differences across major approaches to measuring executive function and self-regulation in early childhood

	Direct assessments	Adult reports	Observational tools
Typical use	Research on specific sub-constructs of EF and SR	Research on overall EF and SR skills and challenges	Research on overall EF and SR behaviors
Examples	*Individual tasks* Hearts and Flowers Marshmallow task Pencil tap *Batteries* NIH toolbox MEFS PSRA	CBQ BRIEF CBCL SDQ	PSRA assessor report RRSM
Empirical evidence	High	High	Low
Conceptual precision	High	Low	Low
Breadth of EF/SR skills covered	Low	Moderate	High
Objectivity	High	Low	Moderate–high
Ecological validity	Low-moderate	High	High
Interpretability	Low	High	High
Scalability	Moderate	High	Low

a growing number of additional behavioral paradigms (see Carlson 2005 and Ponitz et al. 2008 for reviews). For example, the pencil tap task captures children's skills in inhibitory control by asking them to tap once when an assessor taps twice and vice versa (Luria 1966; Diamond and Taylor 1996; Smith-Donald et al. 2007), whereas the backward digit span measures children's working memory by assessing their accuracy in repeating a series of numbers backward (Wechsler 1945). More recently, direct assessments of EF and SR have been digitized, resulting in the emergence of new computer- and tablet-based tasks that provide more precise data on response time and accuracy than their pencil-and-paper counterparts. The Hearts and Flowers task, for example, asks children to tap on the same side of a tablet screen when a heart appears, and to tap on the opposite side when a similarly colored and sized flower appears, thereby challenging both their attention and their inhibitory control (Davidson et al. 2006).

Two primary advantages of direct assessments, perhaps reflected in their popularity as research tools, are their relative objectivity and conceptual precision. Computer-based assessments, in particular, allow for highly standardized implementation and objective scoring procedures, minimizing the introduction of bias from assessors. Additionally, the scores produced by direct assessments tend to be normally distributed, facilitating their use in research and evaluation (e.g., Brod et al. 2017). Direct assessments are also likely better than other approaches at isolating specific EF and SR skills and micro-level processes through systematically manipulating procedures or stimuli. This level of precision is particularly valuable for understanding the interrelations between different sub-constructs of EF and SR, including—for example—the interplay between emotional and cognitive functions (Blair et al. 2005).

Despite their emphasis on conceptual precision, it is important to note that most direct assessments of EF and SR still suffer from some degree of measurement impurity, meaning that task scores may reflect a larger range of skills than they purport to capture. For example, even a relatively straightforward direct assessment like the backward digit span—which, as noted above, is intended to capture working memory—requires children to pay attention and understand the task instructions, as well as to inhibit the prepotent impulse to simply repeat the numbers in the order in which they were originally recited. This impurity may be particularly prevalent during the early childhood period, when EF and SR sub-constructs are still differentiating and other basic skills needed to understand and complete the tasks (e.g., receptive language) are still developing (Miyake et al. 2000; Wiebe et al. 2011; Zelazo et al. 2016). Indeed, factor analytic work has confirmed the difficulty of drawing distinctions between regulatory sub-constructs using direct assessments in early childhood, with several recent studies supporting unidimensional models of EF during this period (Wiebe et al. 2011; Willoughby et al. 2012).

Task impurity may also be stronger in particular subgroups of children, leading to non-random bias in assessment scores. Low performance by shy children, for example, may be falsely attributed to poor EF and SR skills, rather than the social anxiety of interacting with an adult to complete an assessment (Crozier and Perkins 2002). Similarly, children with limited English language skills may be incorrectly assumed to have low EF and SR if they experience difficulty in understanding the task directions or communicating with the assessor (Abedi 2004). These issues can be mitigated to some extent through the implementation of direct assessments using data collectors and settings that are familiar to children.

A related limitation of direct assessments is that their goal of providing conceptual precision often comes at the price of conceptual breadth. Indeed, very few direct assessments offer comprehensive perspectives on EF and SR as unified constructs that incorporate cognitive, emotional, and social processes. This narrow conceptual focus has been hypothesized as one reason why direct assessments often struggle to detect the effects of early childhood interventions, which often target complex, multi-component regulatory behaviors rather than specific, discrete sub-constructs (Zelazo et al. 2016). In an attempt to mitigate conceptual "siloing", several packages of EF and SR tasks have been developed in recent years, including the behaviorally based Preschool Self-Regulation Assessment (PSRA; Smith-Donald et al. 2007), and the computerized National Institutes of Health (NIH) Toolbox (Zelazo et al. 2013) and Executive Function Touch battery (Willoughby and Blair 2016). These batteries combine individual tasks like the pencil tap and snack delay (in the case of the PSRA) or the Dimensional Change Card Sort (DCCS) and Flanker Task (in the case of the NIH Toolbox) to allow a more holistic perspective on EF and SR skills across multiple constructs. Practically, however, the time and resources needed to implement multiple direct assessments is often preclusive, particularly when working with very young children whose attention spans are limited.

Adult Reports

Adult reports use parents, teachers, and other caregivers to report on children's EF and SR behaviors in the context of their everyday environments. Self-reports that are similar in content and structure are also used for older children. Adult reports can take up to 15 min each to complete, although "short forms" or individual subscales often take much less time (e.g., 5 to 10 min). Psychometric evidence on adult reports is largely positive, with many studies showing adequate and consistent evidence of tools' test–retest reliability, internal consistency, factor structure, and invariance (e.g.,

Gioia et al. 1996; Rothbart et al. 2001; Sherman and Brooks 2010; Sulik et al. 2010). Although adult reports have also been found to be sensitive to intervention impacts (e.g., McCoy et al. 2018), they tend to be much less commonly used in preschool-based evaluations than direct assessments.

Modern adult reports of EF and SR have roots in several different literature studies. In particular, temperament researchers have been leveraging parents' reports to capture children's effortful control skills, among other characteristics, for decades (Rothbart 1981). Rothbart's Children's Behavior Questionnaire, for example, captures parents' ratings of preschoolers' inhibitory control, attentional focusing, low-intensity pleasure, and perceptual (in)sensitivity using a seven-point Likert scale ranging from "extremely untrue of your child" to "extremely true of your child" (Rothbart et al. 2001). Other adult reports of EF and SR share similarities with tools from the clinical literature. For example, rather than focusing on EF and SR *skills*, many adult reports are framed to capture regulatory *challenges*, with their psychometric evidence largely focusing on validation against established clinical thresholds (McCandless and O'Laughlin 2007; Sherman and Brooks 2010). For example, one of the most popular adult reports of early childhood EF is the Behavior Rating Inventory of Executive Function-Preschool Version (BRIEF-P; Gioia et al. 1996). Using the BRIEF-P, parents or teachers rate the frequency with which children demonstrate common symptoms of dysregulation (e.g., "forgets what he/she was doing") using a simple three-level response scale of never, sometimes, often. Attention- and hyperactivity-related subscales from more conceptually comprehensive adult reports such as the Strengths and Difficulties Questionnaire (SDQ; Goodman 1997) and the Child Behavior Checklist (CBCL; Achenbach and Edelbrock 1991) have also been used to capture EF and SR difficulties in both clinical and research settings. Importantly, evidence on the psychometric properties of these tools is somewhat mixed relative to other approaches. The SDQ, for example, has shown a number of different factor structures across culturally diverse samples, suggesting that its underlying constructs may be unstable (e.g., Finch et al. 2018; Stone et al. 2010).

Adult reports have several key advantages for use in research. From a practical perspective, they are inexpensive to implement, requiring only an adult's time to complete. Furthermore, because they are rated by individuals who are familiar with children's typical behaviors, adults who are completing these surveys can consider children's behaviors across a range of situations and settings, allowing for a more generalizable perspective on children's EF and SR skills when compared with other contextually constrained approaches. Finally, adult reports typically cover a range of EF and SR sub-constructs in a single scale. This makes them particularly well suited for researchers who are interested in gaining a comprehensive perspective on children's EF and SR across attentional, social, emotional, and behavioral processes, but potentially less appropriate for those hoping to target specific, isolated constructs—particularly those that are not directly observable (e.g., cognition).

Importantly, adult reports are also limited in several ways. One concern related to more clinically focused adult reports (e.g., those targeting attentional or behavioral challenges) is that although they are useful for discriminating children around a clinical threshold of regulatory difficulty, they may be less well equipped to represent the full distribution of children's EF and SR abilities. As such, they may be less useful for identifying children whose EF and SR skills are particularly exemplary, or even within the normative range. Perhaps the most commonly discussed concern regarding adult ratings of child behavior, however, is their subjectivity. In particular, social desirability bias or the pressure of high-stakes testing may lead some adult raters to over-report children's positive behaviors and under-report their negative behaviors in an attempt to appear more favorable (e.g., Eisenberg et al. 1996). Adults may also have a hard time disentangling EF and SR from other information that they know about that child, including their academic or social skills, their tendency to comply with adult instructions, or even their global perception of how well behaved a child might be (Abikoff et al. 1993; Duckworth and Yeager 2015). These sources of subjectivity may partially explain low to moderate observed correlations between parents' and teachers' reports on the same child using the same scale (Mitsis et al. 2000; Sullivan and Riccio 2007; Youngstrom et al. 2000). Finally, adults may have difficulty in accurately reporting children's behaviors across contexts where they don't typically observe them. Indeed, one study found that parents' reports of children's attention, hyperactivity, and impulsivity in school were more strongly related to their own reports of children's behaviors at home than they were of teachers' reports of children's behavior in school (Mitsis et al. 2000).

Observational Tools

A newer- and therefore less well studied and discussed-addition to the EF and SR literature is the *observational tool*, which uses an independent assessor to rate children's skills as they engage in some sort of activity or task. Observational tools take slightly longer than direct assessments and adult reports to administer, as they require both an observation period (which can be as little as a few minutes) as well as the time it takes to fill in the report or rating form itself (which tends to be comparable to an adult report). Because they are relatively less common than direct assessments and adult reports, the psychometric properties and intervention sensitivity of observational tools have been less well studied.

Several examples of observational tools exist in the early childhood EF and SR literature. In particular, the PSRA direct assessment battery includes a supplementary Assessor Report in which the data collector administering the direct assessments systematically rates children's behaviors over the course of the testing period, including the degree to which the child thought and planned before beginning each task and refrained from indiscriminately touching test materials (Smith-Donald et al. 2007). The PSRA Assessor Report has shown evidence for scalar invariance and criterion validity (Daneri et al. 2018) and is also known to be sensitive to early intervention effects (Raver et al. 2011). The Regulation-Related Skills Measure (RRSM; McCoy et al. 2017a) uses a similar set of observed, Likert-scale items to rate preschoolers' and kindergarteners' EF and SR behaviors not during an assessment, but in an everyday classroom environment, including during student-led activities (e.g., free play), teacher-led activities (e.g., read aloud), and transitions between activities (e.g., clean up). Although less well studied, pilot data on a small sample of children using the RRSM suggest high levels of internal consistency, but potential issues with ceiling effects (McCoy et al. 2017a).

Observational tools share similar pros and cons with direct assessments and adult reports. Observational tools tend to be more objective than adult reports due to the fact that their raters typically do not have a prior relationship with the child they are assessing and are therefore less likely to conflate his or her EF and SR skills with other characteristics. Relative to parents or even teachers, experienced observers also have the advantage of a broad comparison group against which to judge a child's behavior, helping them to calibrate their scores and avoiding reference bias (Heine et al. 2002).

At the same time, observational tools are limited in their conceptual precision. Like adult reports, observational assessments can only capture children's observed behavior. They are unable, however, to operationalize the specific underlying skills (e.g., set-shifting, inhibitory control) or attributes (e.g., motivation, mindsets) that may drive this behavior. Second, unlike adult reports, observational approaches can only make inferences about children's behaviors within a particular place and time. As such, their ability to generalize children's EF and SR across contexts is limited. Third, children may intentionally modify their typical behavior when they know that they are being observed, a phenomenon known as the Hawthorne Effect (Diaper 1990). This may lead to an overestimate of children's typical EF and SR behaviors using observational approaches, particularly for children who are sensitive to observer effects.

Applying Executive Function and Self-Regulation Assessments in Real-World Contexts

Given the breadth and diversity of EF and SR assessments available, it is often challenging for researchers, practitioners, policy makers, and clinicians to determine the "best" measurement approach to meet their goals. This dilemma is increasingly salient for stakeholders working in real-world contexts like schools, daycares, or community settings, for which few tools have been explicitly developed, but in which demand for EF and SR assessment is increasing. Indeed, a large number of programs and curricula have been developed over the past several decades to target children's EF and SR development in these everyday environments (Jacob and Parkinson 2015). Furthermore, several recent policy efforts—including the 2015 federal Every Student Succeeds Act (ESSA)—have pushed states and school districts to widen their definitions of young children's success to include valid, reliable, and comparable non-academic metrics of well-being and school readiness (Darling-Hammond et al. 2016; Grant et al. 2017).

Despite this increasing demand, few—if any—widely accepted metrics exist for measuring EF and SR in everyday environments. Indeed, the traditional foci on laboratory-based settings and clinical goals have limited the use of EF and SR assessments for real-world contexts in several ways that are not often discussed in typical reviews of EF and SR measurement. Below I enumerate these limitations, as well as potential guidance for selecting tools that mitigate them.

Ecological Validity

First, many existing approaches to EF and SR assessment lack evidence for their *ecological validity*. Ecological validity refers to a measure's ability to capture processes that are relevant to real-world behaviors and outcomes. In the case of EF and SR, an ecologically valid measure is one whose scores represent the EF and SR skills necessary to succeed in the classroom (or other environment of interest), including the ability to pay attention, wait for one's turn, and avoid calling out answers. Ecological validity is a critical component of field-based, applied research. Tools that lack ecological validity may fail to detect the improvements in everyday behavior that are often targeted through intervention, or to identify children in need of further supports who struggle in certain contexts but not others.

Two standards are typically used for characterizing the ecological validity of a measure. First, verisimilitude is often judged by expert opinion (i.e., face validity) and is

the extent to which the demands being measured by a particular tool reflect the demands of the everyday environment (Chaytor and Schmitter-Edgecombe 2003; Franzen and Wilhelm 1996). In early childhood, an EF or SR measure with high verisimilitude could be based upon a common activity within a preschool classroom (e.g., a memory game, Simon Says) or an observation of behaviors that are aligned with classroom norms (e.g., sitting still, directing attention toward the teacher, raising hands). Second, veridicality is frequently assessed using tests of criterion validity and reflects the degree to which a given tool's scores are predictive of a different metric of real-world functioning (Chaytor and Schmitter-Edgecombe 2003; Franzen and Wilhelm 1996). In the case of preschool-based EF and SR measurement, one might establish veridicality by exploring a measure's correlation with academic performance, social-emotional skills, or other measures of school readiness.

Of the above-described approaches for assessing EF and SR skills in young children, some may lend themselves more than others to establishing ecological validity. As noted above, a large body of academic literature has relied on direct assessments as the primary tools for evaluating the effectiveness of contextually based interventions. Direct assessments are also increasingly being recommended and used in more high-stakes contexts, including as part of kindergarten entry exams and screeners (Ackerman and Friedman-Krauss 2017). Why might the use of direct assessments of EF and SR be problematic in these instances? Most direct assessments are intended for use in quiet, one-on-one testing situations. Although ideal for minimizing distractions and supporting children's performance, these testing situations are not representative of the conditions under which children must regulate their attention, behavior, and emotions on a regular basis. The typical preschool classroom, for example, has been intentionally designed to stimulate children through sights and sounds that engage them in learning and social interaction. Some children who demonstrate strong regulatory capacity in the more sterile, supervised, and scaffolded context of a one-on-one testing environment may therefore struggle to deploy these skills in the more active, distracting, and emotionally charged classroom context. Other children, however, may find it easier to demonstrate EF and SR skills when surrounded by a familiar setting with established incentives.

In this sense, traditional direct assessments may incorrectly estimate many children's day-to-day EF and SR behaviors by neglecting to account for real-world factors that challenge and support the deployment of multiple EF and SR skills (Duckworth and Yeager 2015; McClelland and Cameron 2012). This theory is supported by research examining the distinction between "hot" and "cool" EF. Hot EF refers to the deployment of attention, inhibition, and working memory skills in the context of "motivationally and emotionally significant situations", whereas cool EF refers to the use of these skills in the absence of such demands (Zelazo and Carlson 2012).

Research has shown that young children's performance on EF direct assessments tends to differ based on whether the task is intended to be "hot" versus "cool". On the one hand, prior evidence has found that EF performance may be diminished in the context of enhanced motivational and emotional influences (e.g., when using a reward of an enticing jelly bean versus an unappealing abstract symbol; Carlson 2005), suggesting that hot EF may require a more advanced or complex set of skills relative to cool EF (Zelazo and Carlson 2012). Evidence from neuroscience supports this hypothesis, showing that motivational demands activate additional parts of the brain above and beyond those required for cool EF (Zelazo and Carlson 2012). On the other hand, an additional set of studies has shown that hot tasks may actually facilitate performance for children when they introduce positive or rewarding conditions. For example, Qu and Zelazo (2007) found improved performance on the DCCS when children were asked to sort faces showing happy—but not sad or neutral-emotions, and speculated that these differences may have been driven by increased dopamine levels in the area of the brain that supports cognitive flexibility. Collectively, this evidence implies that hot EF assessments are more likely than their cool counterparts to be ecologically valid, as they require children to recruit a number of regulatory processes in the context of external rewards and motivators in order to appropriately meet the demands of the task.

Although hot EF tasks may be more ecologically valid than cool ones, it is unclear the degree to which direct assessments can ever truly mimic the complexity of "hot" situations faced by children in everyday environments. The introduction of a desirable stimulus such as a marshmallow or jellybean may successfully evoke a motivational or emotional response from the child being assessed, but it will not represent the full range of possible challenges, temptations, or supports faced by that child under typical circumstances. Indeed, early childhood classrooms are full of distractions, demands, and incentives (e.g., noise levels, expectations and reinforcements from teachers) that are impossible to replicate in a systematic way due to their variability and differential salience to particular children (Duckworth and Yeager 2015).

Given this complexity, it is clear that assessments of EF and SR administered outside of real-world contexts may fail to achieve ecological validity in that they are unable to replicate the demands and expectations of children's environments (i.e., they are unable to establish verisimilitude). As such, researchers looking to maximize the ecological validity of their measurement tools may want to prioritize adult reports or observational tools over more traditional direct

assessment approaches. Although limited in other ways (see above), tools like the BRIEF-P and RRSM are designed to comprehensively assess children's EF and SR skills *in context*, during the day-to-day interactions, distractions, temptations, and demands that shape these skills with time. The BRIEF-P, for example, is explicitly intended for use in "everyday contexts" and includes reference to behaviors like getting out of a seat and difficulties with concentrating on chores and schoolwork (Isquith et al. 2004). These skill types are both consistent with the goal of verisimilitude and more proximal to the concrete behavioral strategies encouraged by most EF and SR interventions (e.g., stop, breathe, think; Dusenbury et al. 2015). As such, these approaches may maximize researchers' odds of detecting impacts of behavioral interventions that may otherwise go undetected by more distal, micro-level direct assessment tools.

At the same time, new research from Obradović et al. (2018) using an older sample of elementary school students suggests that subtle modifications to the implementation of direct assessments can also support improved veridicality of these tools, while retaining their other advantages in objectivity and conceptual precision. These researchers compared students' EF performance using the Hearts and Flowers task administered in a traditional one-on-one setting versus the same task administered in a classroom-based, group setting. They found that both the individual- and group-based administration produced scores that were highly correlated with one another, equally reliable, and similarly predictive of teacher EF ratings and standardized test scores. However, only the group-based EF score was predictive of gains in students' academic achievement, suggesting that performance on the Hearts and Flowers in the context of distractions from peers and demands from the classroom context captures unique and relevant skills "above and beyond" those captured in a one-on-one setting. Importantly, Obradović et al. also note that the group-based approach was easier and faster to administer than the one-on-one format and therefore could represent both a conceptual and pragmatic improvement over traditional direct assessment paradigms. Future research is needed to determine whether group-based administration may be a similarly useful- and practical-approach for assessing EF and SR in younger children.

It is important to note that ecological validity comes with specific trade-offs. By definition, scores on ecologically valid EF and SR assessments capture aspects of both children's latent EF and SR skills, *as well as* the demands and supports provided by the context in which they were measured. Although all EF and SR assessments are likely to conflate ability and contextual demands to some extent (see, for example, recent research demonstrating the conceptual impurity of the marshmallow task; Frankenhuis et al. 2016; Kidd et al. 2013; Lamm et al. 2018; Watts et al. 2018), researchers or stakeholders interested in capturing "pure" or optimal representations of children's EF and SR abilities should prioritize standardization over ecological validity when selecting an assessment metric.

Furthermore, when interpreting scores from ecologically embedded assessments (e.g., adult reports, observational tools), it is important to consider the extent to which children's scores may have differed under alternative situational supports or constraints. The RRSM observational tool, for example, recommends that users score each child's EF and SR behaviors within at least three different common classroom contexts (i.e., a teacher-led activity, a student-led activity, and a transition) in order to quantify the variability in his or her skills across different situational demands. Results of pilot work suggest that as much as half of the variation in individuals' RRSM scores is attributable to these different situational conditions, whereas the other half is attributable to more stable skill levels of the child (McCoy et al. 2017b). These findings reinforce the importance of assessing EF and SR skills within the exact context of interest to a particular study, rather than assuming that skill deployment is consistent across situations.

Interpretability

In addition to ecological validity, a second consideration for applied stakeholders working in classrooms and other real-world contexts is the *interpretability* of the scores produced by EF and SR measures. Interpretable scores can be easily linked with concrete behaviors, making it easy for stakeholders to identify their significance and use them for decision-making. More clinically oriented adult reports, for example, often come with cutoffs for clinically significant EF and SR problems that were validated using norm-referenced populations and in-depth criterion measures (e.g., clinical interviews; McCandless and O'Laughlin 2007; Sherman and Brooks 2010). As such, these tools are quite useful for informing decisions regarding whether or not to refer a child for additional screening or services by comparing his or her performance to other children of the same age (Zelazo et al. 2016).

In addition to identifying individual children at risk of EF or SR challenges, stakeholders may also be interested in using data from EF and SR measures for tracking individual growth over time, or for informing action on a different level. Teachers may be interested, for example, in monitoring their students' progress in EF and SR development over the course of a school year, or in identifying individual children who are excelling in EF and SR and in need of additional resources to support further growth. At a broader level, policy makers or school officials may want to conduct a needs assessment to understand more about the specific EF and SR challenges facing their populations of interest, helping them to develop more informed interventions.

Furthermore, as research journals increasingly adopt standards for reporting practical—rather than statistical—significance, applied researchers may become more and more invested in explicitly linking their findings to real-world EF and SR benchmarks.

Certain EF and SR tools may be more versus less well positioned to achieve these various goals. Most adult reports and observational tools use Likert-type scales with frequency anchors (e.g., 0 = never, 1 = sometimes, 2 = often), making it easy to calculate average scale scores with inherent meaning (e.g., where a mean score of 0.2 would suggest that a child, on average, very rarely exhibits the measured behaviors) that can be compared across time and age. At the same time, the comprehensive nature of these tools means that they are more well suited for identifying global strengths and weaknesses, rather than picking up on the specific skills (e.g., inhibitory control, working memory) in which children need more versus less support. Furthermore, as noted above, clinically oriented adult reports tend to focus on children who are struggling in their EF and SR development and may therefore not be particularly adept at discriminating between children who are doing well or excelling at these skills.

Direct assessments, on the other hand, are intended to give more targeted information on specific EF and SR skills across the full normal curve of EF and SR ability. As such, when multiple direct assessments are administered to an individual or sample (e.g., using a battery), the scores provided can help stakeholders to understand which particular skills are in need of further intervention and support, for what range of ability levels. Despite these advantages, the scores produced by direct assessments tend to be more abstract than those provided by adult report and observational tools. In particular, direct assessments often examine an individual's performance on a task using average response time or percent of trials scored as "correct", which are difficult to directly link with an individual's everyday behavior. Furthermore, researchers often standardize direct assessment scores either within their samples (e.g., using a *z*-score; Smith-Donald et al. 2007) or against external population norms (e.g., Carlson and Zelazo 2014) to allow comparison of an individual's or group's EF and SR skill levels relative to others'. This standardization can be fruitful for stakeholders interested in knowing how their children's skill levels compare to their peers' (e.g., using a percentile rank), but is less useful for describing children's *absolute* ability levels, or what particular regulation-related behaviors they have and have not yet mastered.

Overall, it is clear that each approach to EF and SR assessment has a unique set of strengths with regard to its interpretability. At the same time, it is important to recognize that few—if any—EF and SR measures show promise for guiding actual practice or differentiated instruction. In other words, although EF and SR tools may be useful in providing information on *who* is in need of additional support in a given area, they are generally not intended to guide *what* those supports should look like or *how* they should be implemented. Future research is needed to generate formative tools that can help teachers and other stakeholders to identify the practices that are most closely linked with EF and SR outcomes, as well as to provide responsive guidance on the types of activities that might facilitate skill development in a given area.

Scalability

Finally, an increasing concern of stakeholders interested in applied EF and SR assessment is the *scalability* of various measurement approaches. As noted above, increased public understanding of the importance of EF and SR skill development has led to a growth in demand for the monitoring of these skills as part of policy and standardized testing efforts (Ackerman and Friedman-Krauss 2017; Darling-Hammond et al. 2016; Grant et al. 2017). To meet this demand, a number of approaches have been used to provide quick and inexpensive data at a large scale.

First, direct assessments are currently being considered for inclusion in several state-level standardized testing batteries (e.g., the Texas Kindergarten Entry Assessment) and have been used by private companies (e.g., Reflection Sciences, LLC) to provide data on EF and SR skills to a variety of stakeholders. Computerized direct assessments offer particular promise for scale, as they can be implemented relatively quickly, either with the support of a data collector (e.g., within the classroom setting) or without (e.g., virtually, using a family computer or parent's cell phone). At the same time, these assessments assume access to technological devices (e.g., laptops, tablets) that may not reliably exist, particularly in low-resource settings. Furthermore, most direct assessments have not been tested or validated for non-English-speaking populations, making it difficult to deploy them in multi-lingual contexts (McClelland and Cameron 2012). Large-scale administration of direct assessments also increases the likelihood of inconsistent implementation (e.g., where some children receive more support to complete the task than others), which can systematically bias children's scores (Duckworth and Yeager 2015).

Adult reports are also quite popular for use at scale, and an increasing number of accountability and monitoring systems are incorporating adult-reported data on EF and SR. Canada's teacher-reported Early Development Instrument (EDI), for example, includes a set of questions about children's hyperactive and inattentive behavior, and has been implemented for more than 1.1 million children around the world (Janus and Offord 2007). Several traditional adult reports (e.g., the BRIEF-P) have also published short forms, which select the most informative items from a longer item

pool and therefore make assessments more practical and efficient (LeJeune et al. 2010). Although likely the most scalable approach to EF and SR assessment in early childhood, adult reports come with some logistical challenges, including difficulty accessing parents (which may lead to missing data for particular groups of children) or over-burdening teachers by asking them to report on multiple children (which may lead to low-quality data). Furthermore, they are not advised for use in high-stakes testing efforts due to strong potential for bias, particularly when teachers' ratings of children are tied to their own pay or retention (Regenstein et al. 2017).

Finally, observational approaches have been infrequently used at scale, likely due to the burden they place on data collectors. Nevertheless, as standardized assessments and other forms of testing become increasingly popular in applied settings (Regenstein et al. 2017; Stedron and Berger 2010), new opportunities may develop for building in observational reports of children's behavior during testing.

Future Directions

Although the field has made tremendous progress over the past several decades in developing new approaches for capturing young children's EF and SR skills, a number of open questions remain. Addressing these questions will be critical for informing the relevance of various EF and SR assessments for capturing children's skills in their classrooms and other environments in which they spend their time. First, additional research is needed to explicitly test the degree to which different measures of children's EF and SR skills are sensitive to and potentially dependent on the contextual demands that children face on a daily basis. As noted above, preliminary evidence from the RRSM and group-based Hearts and Flowers suggests that children's EF and SR scores may vary depending on the situations in which assessments are implemented (e.g., in teacher-directed versus student directed contexts, in individual versus group-based settings). Future research quantifying and predicting this variability will complement existing work on the verisimilitude and veridicality of EF and SR measures to generate greater evidence on their ecological sensitivity and, by extension, validity.

Second, further research is needed to understand the extent to which various EF and SR assessments may be more or less sensitive to intervention. As noted above, direct assessments have been used extensively as outcome measures in the context of EF and SR evaluation studies. Although these tools tend to be objective and conceptually precise (two important advantages in the context of research), they may not always align well with the conceptual targets of preschool-based interventions (Jones et al. 2017). Moving forward, researchers should consider including adult reported and observational tools as intervention outcomes, as these tools may be more well suited to capturing the ecologically based behaviors (e.g., focusing attention, regulating impulses and emotions) often targeted in early childhood EF and SR programming. Given their potential for scalability and existing use in accountability and monitoring systems, exploring the responsiveness of adult reports like the EDI to intervention and policy is a particularly important need for the field.

Finally, work is also needed to develop more formative tools that can either replace or complement existing approaches to EF and SR assessment. In particular, there is a pressing need for actionable guidance that teachers, administrators, and clinicians can use to change their practices in response to children's assessed EF and SR skill levels. A handful of newer tools tool have been designed for such purposes. The Teaching Strategies GOLD®, for example, use early childhood classroom teachers as assessors and includes a set of items targeting regulatory skills as part of its broader social-emotional domain (Lambert et al. 2015). Nevertheless, formative assessments that explicitly "unpack" and respond to the various subcomponents of EF and SR remain lacking. Developing such tools will be instrumental for supporting stakeholders in providing the types of activities, practices, and individualized instruction that are known to build children's EF and SR (Jones et al. 2017).

Conclusions

Although early childhood researchers have made great strides in measuring children's foundational EF and SR skills, the diversity of EF and SR measures available to stakeholders often raises confusion regarding which approach to use for which particular purpose. Research has shown that correlations among different EF and SR measurement approaches tend to be low, with one review of 20 studies showing a median correlation between direct assessments and adult reports of just $r = 0.19$ (Toplak et al. 2013). Indeed, it is clear from the evidence reviewed above that the information gleaned from these different approaches to EF and SR assessment is fundamentally different and, as a result, non-comparable.

Direct assessments, adult reports, and observational approaches each demonstrate a unique set of trade-offs with regard to their empirical properties, conceptual breadth and depth, and practical constraints. Researchers interested in measuring young children's EF and SR skills in real-world environments (e.g., classrooms) must grapple with these tradeoffs, making informed decisions to minimize bias while maximizing the likelihood of achieving their particular goals. Overall, the evidence reviewed in this manuscript suggests that one-on-one direct assessments, although relatively

conceptually precise and free of bias, may be limited in their ability to provide interpretable scores, as well as their capacity to capture the complex confluence of "hot" and "cool" skills that children need to succeed in the real world. Adult reports and observational tools, on the other hand, hold promise in terms of their ecological validity and their ability to directly translate into meaningful, interpretable metrics, yet are limited in a number of other ways. In particular, observational tools are resource intensive to implement, therefore limiting their scalability.

As the demand for applied measures of EF and SR continues to grow, future research is needed to better understand the advantages and disadvantages of assessing these constructs in real-world environments. Research is particularly needed to support the development and adaptation of EF and SR measures that are interpretable, actionable, and scalable for diverse populations. Such work will not only facilitate more in-depth, ecologically informed research on basic EF and SR processes, but will also lay the foundation for more effective intervention, practice, and policy in the early childhood period.

Compliance with Ethical Standards

Conflict of interest The author declares that she has no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by the author.

References

Abedi, J. (2004). The no child left behind act and English language learners: Assessment and accountability issues. *Educational Researcher, 33*(1), 4–14.

Abikoff, H., Courtney, M., Pelham, W. E., & Koplewicz, H. S. (1993). Teachers' ratings of disruptive behaviors: The influence of halo effects. *Journal of Abnormal Child Psychology, 21*(5), 519–533.

Achenbach, T. M., & Edelbrock, C. (1991). Child behavior checklist. *Burlington (Vt), 7*.

Ackerman, D. J., & Friedman-Krauss, A. H. (2017). Preschoolers' executive function: Importance, contributors, research needs and assessment options. *ETS Research Report Series, 2017*(1), 1–24.

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development, 78*(2), 647–663.

Blair, C., Zelazo, P. D., & Greenberg, M. T. (2005). The measurement of executive function in early childhood. *Developmental Neuropsychology, 28*(2), 561.

Brod, G., Bunge, S. A., & Shing, Y. L. (2017). Does one year of schooling improve children's cognitive control and alter associated brain activation? *Psychological Science, 28*(7), 967–978.

Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology, 28*(2), 595–616.

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

Chaytor, N., & Schmitter-Edgecombe, M. (2003). The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychology Review, 13*(4), 181–197.

Crozier, W. R., & Perkins, P. (2002). Shyness as a factor when assessing children. *Educational Psychology in Practice, 18*(3), 239–244.

Daneri, M. P., Sulik, M. J., Raver, C. C., & Morris, P. A. (2018). Observers' reports of self-regulation: Measurement invariance across sex, low-income status, and race/ethnicity. *Journal of Applied Developmental Psychology, 55*, 14–23.

Darling-Hammond, L., Bae, S., Cook-Harvey, C. M., Lam, L., Mercer, C., Podolsky, A., & Stosich, E. L. (2016). *Pathways to new accountability through the Every Student Succeeds Act*. Palo Alto: Learning Policy Institute.

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia, 44*(11), 2037–2078.

Diamond, A., & Lee, K. (2011). Interventions shown to aid executive function development in children 4 to 12 years old. *Science, 333*(6045), 959–964.

Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the abilities to remember what I said and to "Do as I say, not as I do". *Developmental Psychobiology, 29*(4), 315–334.

Diaper, G. (1990). The Hawthorne effect: A fresh examination. *Educational studies, 16*(3), 261–267.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44*(4), 237–251.

Dusenbury, L., Calin, S., Domitrovich, C., & Weissberg, R. P. (2015). *What does evidence-based instruction in social and emotional learning actually look like in practice? A brief on findings from CASEL's program reviews* (p. 3). Chicago, IL: Collaborative for Academic, Social, and Emotional Learning.

Eisenberg, N., Fabes, R. A., & Murphy, B. C. (1996). Parents' reactions to children's negative emotions: Relations to children's social competence and comforting behavior. *Child Development, 67*(5), 2227–2247.

Finch, J. E., Yousafzai, A. K., Rasheed, M., & Obradović, J. (2018). Measuring and understanding social-emotional behaviors in preschoolers from rural Pakistan. *PLoS ONE, 13*(11), e0207807.

Fox, E., & Riconscente, M. (2008). Metacognition and self-regulation in James, Piaget, and Vygotsky. *Educational Psychology Review, 20*(4), 373–389.

Frankenhuis, W. E., Panchanathan, K., & Nettle, D. (2016). Cognition in harsh and unpredictable environments. *Current Opinion in Psychology, 7*, 76–80.

Franzen, M. D., & Wilhelm, K. L. (1996). Conceptual foundations of ecological validity in neuropsychological assessment.

Gioia, G. A., Andrwes, K., & Isquith, P. K. (1996). *Behavior rating inventory of executive function-preschool version (BRIEF-P)*. Odessa: Psychological Assessment Resources.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, 38*(5), 581–586.

Grant, S., Hamilton, L. S., Wrabel, S. L., Gomez, C. J., Whitaker, A. … Ramos, A. (2017). *How the Every Student Succeeds Act can support social and emotional learning*. Santa Monica: RAND Corporation.

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. *Journal of Personality and Social Psychology, 82*(6), 903.

Isquith, P. K., Gioia, G. A., & Espy, K. A. (2004). Executive function in preschool children: Examination through everyday behavior. *Developmental Neuropsychology, 26*(1), 403–422.

Jacob, R., & Parkinson, J. (2015). The potential for school-based interventions that target executive function to improve academic achievement: A review. *Review of Educational Research, 85*(4), 512–552.

Janus, M., & Offord, D. R. (2007). Development and psychometric properties of the Early Development Instrument (EDI): A measure of children's school readiness. *Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement, 39*(1), 1.

Jones, S. M., Bailey, R., Barnes, S. P., & Partee, A. (2016). Executive function mapping project executive summary: Untangling the terms and skills related to executive function and self-regulation in early childhood. In *OPRE Report # 2016-88*. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Jones, S. M., Barnes, S. P., Bailey, R., & Doolittle, E. J. (2017). Promoting social and emotional competencies in elementary school. *The Future of Children, 27*, 49–72.

Kidd, C., Palmeri, H., & Aslin, R. N. (2013). Rational snacking: Young children's decision-making on the marshmallow task is moderated by beliefs about environmental reliability. *Cognition, 126*(1), 109–114.

Lambert, R. G., Kim, D. H., & Burts, D. C. (2015). The measurement properties of the Teaching Strategies GOLD® assessment system. *Early Childhood Research Quarterly, 33*, 49–63.

Lamm, B., Keller, H., Teiser, J., Gudi, H., Yovsi, R. D., Freitag, C., Vöhringer, I. (2018). Waiting for the second treat: Developing culture-specific modes of self-regulation. *Child Development, 89*(3), e261–e277.

LeJeune, B., Beebe, D., Noll, J., Kenealy, L., Isquith, P., & Gioia, G. (2010). Psychometric support for an abbreviated version of the Behavior Rating Inventory of Executive Function (BRIEF) Parent Form. *Child Neuropsychology, 16*(2), 182–201.

Luria, A. (1966). *Higher cortical functions in man*. New York: Basic Books.

McCandless, S., & O'Laughlin, L. (2007). The clinical utility of the Behavior Rating Inventory of Executive Function (BRIEF) in the diagnosis of ADHD. *Journal of Attention Disorders, 10*(4), 381–389.

McClelland, M. M., & Cameron, C. E. (2012). Self-regulation in early childhood: Improving conceptual clarity and developing ecologically valid measures. *Child Development Perspectives, 6*, 136–142.

McCoy, D. C. (2013). Early violence exposure and self-regulatory development: A bioecological systems perspective. *Human Development, 56*(4), 254–273.

McCoy, D. C., Jones, S. M., Hemenway, A., Koepp, A., & Wilder-Smith, O. (2017a, November). *An Observational Measure of Regulation-Related Skills in the Early Childhood Classroom Setting*. In Paper presented at the meeting of the Association for Public Policy Analysis & Management, Chicago, IL.

McCoy, D. C., Jones, S., Leong, D., Bodrova, E., Wilder-Smith, B., & Koepp, A. (2017b, April). *An Observational Measure of Regulation-Related Skills in the Early Childhood Classroom Setting*. In Paper presented at the meeting of the Society for Research in Child Development, Austin, TX.

McCoy, D. C., Jones, S., Roy, A., & Raver, C. C. (2018). Classifying social-emotional trajectories through elementary school: Impacts of the Chicago School Readiness Project. *Developmental Psychology, 54*, 772–787.

Mischel, W., Ebbesen, E. B., & Raskoff Zeiss, A. (1972). Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology, 21*(2), 204.

Mitsis, E. M., McKay, K. E., Schulz, K. P., Newcorn, J. H., & Halperin, J. M. (2000). Parent–teacher concordance for DSM-IV attention-deficit/hyperactivity disorder in a clinic-referred sample. *Journal of the American Academy of Child & Adolescent Psychiatry, 39*(3), 308–313.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49–100.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., …, Sears, M. R. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences, 108*(7), 2693–2698.

Morris, P., Mattera, S., Castells, N., Bangser, M., Bierman, K., & Raver, C. C. (2014). Impact findings from the Head Start CARES demonstration: National evaluation of three approaches to improving preschoolers' social and emotional competence.

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

Obradović, J., Sulik, M. J., Finch, J. E., & Tirado-Strayer, N. (2018). Assessing students' executive functions in the classroom: Validating a scalable group-based procedure. *Journal of Applied Developmental Psychology, 55*, 4–13.

Pandey, A., Hale, D., Das, S., Goddings, A. L., Blakemore, S. J., & Viner, R. (2017). Effectiveness of universal self-regulation-based interventions to improve self-regulation, and effects on distant health and social outcomes in children and adolescents: A systematic review and meta-analysis. *The Lancet, 390*, S66.

Ponitz, C. E. C., McClelland, M. M., Jewkes, A. M., Connor, C. M., Farris, C. L., & Morrison, F. J. (2008). Touch your toes! Developing a direct measure of behavioral regulation in early childhood. *Early Childhood Research Quarterly, 23*(2), 141–158.

Qu, L., & Zelazo, P. D. (2007). The facilitative effect of positive stimuli on 3-year-olds' flexible rule use. *Cognitive Development, 22*(4), 456–473.

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Bub, K., & Pressler, E. (2011). CSRP's impact on low-income preschoolers' preacademic skills: self-regulation as a mediating mechanism. *Child Development, 82*(1), 362–378.

Regenstein, E., Connors, M., Romero-Jurado, R. I. O., & Weiner, J. (2017). *Uses and misuses of kindergarten readiness assessment results*. Chicago: The Ounce of Prevention Fund.

Rothbart, M. K. (1981). Measurement of temperament in infancy. *Child Development, 52*, 569–578.

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at three to seven years: The Children's Behavior Questionnaire. *Child Development, 72*(5), 1394–1408.

Schlam, T. R., Wilson, N. L., Shoda, Y., Mischel, W., & Ayduk, O. (2013). Preschoolers' delay of gratification predicts their body mass 30 years later. *The Journal of Pediatrics, 162*(1), 90–93.

Sherman, E. M., & Brooks, B. L. (2010). Behavior rating inventory of executive function–preschool version (BRIEF-P): Test review and clinical guidelines for use. *Child Neuropsychology, 16*(5), 503–519.

Smith-Donald, R., Raver, C. C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-regulation Assessment (PSRA) for field-based research. *Early Childhood Research Quarterly, 22*(2), 173–187.

Stedron, J. M., & Berger, A. (2010). *NCSL technical report: State approaches to school readiness assessment*. Washington, D.C.: National Conference of State Legislatures.

Stone, L. L., Otten, R., Engels, R. C., Vermulst, A. A., & Janssens, J. M. (2010). Psychometric properties of the parent and teacher versions of the strengths and difficulties questionnaire for 4- to 12-year-olds: A review. *Clinical Child and Family Psychology Review, 13*(3), 254–274.

Sulik, M. J., Huerta, S., Zerr, A. A., Eisenberg, N., Spinrad, T. L., Valiente, C., … Edwards, A. (2010). The factor structure of effortful control and measurement invariance across ethnicity and sex in a high-risk sample. *Journal of Psychopathology and Behavioral Assessment, 32*(1), 8–22.

Sullivan, J. R., & Riccio, C. A. (2007). Diagnostic group differences in parent and teacher ratings on the BRIEF and Conners' scales. *Journal of Attention Disorders, 11*(3), 398–406.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Practitioner review: Do performance-based measures and ratings of executive function assess the same construct? *Journal of Child Psychology and Psychiatry, 54*(2), 131–143.

Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science*.

Wechsler, D. (1945). A standardized memory scale for clinical use. *The Journal of Psychology, 19*(1), 87–95.

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development, 84*(6), 2112–2130.

Wiebe, S. A., Sheffield, T., Nelson, J. M., Clark, C. A., Chevalier, N., & Espy, K. A. (2011). The structure of executive function in 3-year-olds. *Journal of Experimental Child Psychology, 108*(3), 436–452.

Willoughby, M. T., & Blair, C. B. (2016). Longitudinal measurement of executive function in preschoolers. In J. A. Griffin, P. McCardle & L. S. Fruend (Eds.), *Executive function in preschool-age children: Integrating measurement, neurodevelopment, and translational research* (pp. 91–113). Washington, DC: American Psychological Association.

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

Youngstrom, E., Loeber, R., & Stouthamer-Loeber, M. (2000). Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *Journal of Consulting and Clinical Psychology, 68*(6), 1038.

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., & Weintraub, S. (2013). II. NIH Toolbox Cognition Battery (CB): Measuring executive function and attention. *Monographs of the Society for Research in Child Development, 78*(4), 16–33.

Zelazo, P. D., Blair, C. B., & Willoughby, M. T. (2016). *Executive function: Implications for education. NCER 2017–2000*. Washington, DC: National Center for Education Research.

Zelazo, P. D., & Carlson, S. M. (2012). Hot and cool executive function in childhood and adolescence: Development and plasticity. *Child Development Perspectives, 6*(4), 354–360.