The Generalizability of Two Rating Scales of Executive Function Across Parent and Self-Reports in a School-based Adolescent Sample

Leah Singh

Follow this and additional works at: https://digitalcommons.memphis.edu/etd

Recommended Citation

This Dissertation is brought to you for free and open access by University of Memphis Digital Commons. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of University of Memphis Digital Commons. For more information, please contact khggerty@memphis.edu.
THE GENERALIZABILITY OF TWO RATING SCALES OF EXECUTIVE FUNCTION ACROSS PARENT AND SELF-REPORT IN A SCHOOL-BASED ADOLESCENT SAMPLE

by

Leah Jai Singh

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

Major: Psychology

The University of Memphis

August 2018
Abstract

This study examined the reliability, validity, and generalizability of test scores obtained from parents’ and adolescents’ reports across two rating scales of executive function. Forty-two adolescents ages 12 to 17 enrolled in a rural public school and one of their parents participated in the study. For each adolescent-parent dyad, 2 total scores of executive function and 6 subscale scores hypothesized to measure the same theoretical constructs were calculated and scores were submitted to several Generalizability theory analyses to evaluate the instrument effect, rater effect, dimension effect, and all interactions on total scores and subscale scores. The resulting dependability coefficients were markedly low (i.e., .16 to .69) and much lower than expected given prior research and the corresponding Pearson correlations evaluating these facets in isolation. The subject-by-rater interaction contributed the largest proportion of variance in test scores (i.e., 30% to 33%). Surprisingly, the dimension effect and its interactions contributed little variation in test scores, suggesting that the construct of executive function is unidimensional. Results indicated that most of the subscale score variance was due to the inclusion of the rater facet, and particularly to the adolescent raters. The results from this study emphasize the importance of using extreme caution when generalizing scores of executive function across different instruments, raters, and specific executive functions.

Keywords: executive function rating scales, construct validity, Generalizability theory, dependability coefficients
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Executive Functioning in Children and Adolescents</td>
<td>2</td>
</tr>
<tr>
<td>The Ambiguity of Executive Functioning</td>
<td>3</td>
</tr>
<tr>
<td>Measuring Executive Functioning in Children and Adolescents</td>
<td>5</td>
</tr>
<tr>
<td>Performance-based Tests</td>
<td>5</td>
</tr>
<tr>
<td>Behavior Rating Scales</td>
<td>6</td>
</tr>
<tr>
<td>Psychometric Standards and Behavior Rating Scales Measuring Executive Functions</td>
<td>9</td>
</tr>
<tr>
<td>Source and Setting Variance</td>
<td>10</td>
</tr>
<tr>
<td>Instrument Variance</td>
<td>16</td>
</tr>
<tr>
<td>Dimension Variance</td>
<td>19</td>
</tr>
<tr>
<td>Purpose of the Study</td>
<td>23</td>
</tr>
<tr>
<td>2 Method</td>
<td>26</td>
</tr>
<tr>
<td>Participants</td>
<td>26</td>
</tr>
<tr>
<td>Measures</td>
<td>27</td>
</tr>
<tr>
<td>Behavior Rating Scale of Executive Function, Second Edition</td>
<td>27</td>
</tr>
<tr>
<td>Comprehensive Executive Function Inventory</td>
<td>32</td>
</tr>
<tr>
<td>Procedures</td>
<td>37</td>
</tr>
<tr>
<td>Recruitment and Selection of Participants</td>
<td>37</td>
</tr>
<tr>
<td>Post-consent Contact and Scheduling of Rating Scale Administration</td>
<td>37</td>
</tr>
<tr>
<td>Parent Administration</td>
<td>38</td>
</tr>
<tr>
<td>Adolescent Administration</td>
<td>40</td>
</tr>
<tr>
<td>3 Results</td>
<td>43</td>
</tr>
<tr>
<td>Data Screening and Tests of Assumptions</td>
<td>43</td>
</tr>
<tr>
<td>Inter-rater Reliability and Mean Differences across Raters</td>
<td>45</td>
</tr>
<tr>
<td>Convergent Validity and Mean Differences across Instruments</td>
<td>50</td>
</tr>
<tr>
<td>Generalizability Theory Analyses</td>
<td>58</td>
</tr>
<tr>
<td>Partial Models</td>
<td>63</td>
</tr>
<tr>
<td>4 Discussion</td>
<td>69</td>
</tr>
<tr>
<td>Dependability of Scores for Executive Function Rating Scales</td>
<td>70</td>
</tr>
<tr>
<td>Instrument Effects</td>
<td>71</td>
</tr>
<tr>
<td>Dimension Effects</td>
<td>78</td>
</tr>
<tr>
<td>Rater Effects</td>
<td>80</td>
</tr>
<tr>
<td>Limitations and Directions for Future Research</td>
<td>85</td>
</tr>
<tr>
<td>Implications for Practice</td>
<td>88</td>
</tr>
<tr>
<td>References</td>
<td>91</td>
</tr>
</tbody>
</table>
Appendices

A. 105
B. 106
C. 107
D. 110
E. 112
F. 113
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Description of BRIEF2 Subscales and Global Executive Composite</td>
<td>28</td>
</tr>
<tr>
<td>2.</td>
<td>Description of CEFI Subscales and Full Scale</td>
<td>33</td>
</tr>
<tr>
<td>3.</td>
<td>Descriptive Statistics by Instrument across Rater, and Inter-Rater Reliability Correlations for Total Scores and Subscale Scores</td>
<td>48</td>
</tr>
<tr>
<td>4.</td>
<td>Intercorrelations across Instrument within Rater for Adolescents</td>
<td>54</td>
</tr>
<tr>
<td>5.</td>
<td>Intercorrelations across Instrument within Rater for Parents</td>
<td>56</td>
</tr>
<tr>
<td>6.</td>
<td>Variance Component Estimates and Absolute Dependability Coefficients by Score Comparison</td>
<td>59</td>
</tr>
<tr>
<td>7.</td>
<td>Variance Component Estimates and Absolute Dependability Coefficients by Subscales (Including the CEFI Planning Subscale)</td>
<td>61</td>
</tr>
<tr>
<td>8.</td>
<td>Variance Component Estimates and Absolute Dependability Coefficients by Subscales (Including the CEFI Organization Subscale)</td>
<td>63</td>
</tr>
<tr>
<td>9.</td>
<td>Variance Component Estimates and Absolute Dependability Coefficients by Subscales for the Parent Rater and Including the CEFI Planning Subscale</td>
<td>65</td>
</tr>
<tr>
<td>10.</td>
<td>Variance Component Estimates and Absolute Dependability Coefficients by Subscales for the Adolescent Rater and Including the CEFI Planning Subscale</td>
<td>66</td>
</tr>
<tr>
<td>11.</td>
<td>Variance Component Estimates and Absolute Dependability Coefficients by Subscales for the Parent Rater and Including the CEFI Organization Subscale</td>
<td>67</td>
</tr>
<tr>
<td>12.</td>
<td>Variance Component Estimates and Absolute Dependability Coefficients by Subscales for the Adolescent Rater and Including the CEFI Organization Subscale</td>
<td>68</td>
</tr>
</tbody>
</table>
The Generalizability of Executive Functioning Ratings Across Parent and Self-Reports in a School-based Adolescent Sample

Over the past few decades, the methods and techniques utilized in neuropsychological evaluations have significantly advanced across the field of psychology and particularly, the fields of clinical child psychology, pediatric psychology and neuropsychology, and school psychology (Miller, 2009; Williams & Boll, 1997). Neuropsychological evaluations utilize a multi-source, multi-method approach, which includes narrow band neuropsychology measures as well as broader measures of intelligence, academic achievement, and informant rating scales to evaluate the severity of cognitive dysfunction (Miller, 2009). The aim of neuropsychological assessment is to examine brain-behavior relationships associated with executive functioning in order to understand why and by which process cognitive and behavioral dysfunction exist (Benton, Hamsher, Varney, & Spreen, 1983). Executive functioning represents a broad construct that refers to a system of cognitive processes that facilitates complex abilities such as verbal and perceptual reasoning, memory, attentional control, cognitive efficiency, and cognitive flexibility (Harvey, 2012; Toplak et al., 2013). Furthermore, research indicates that executive functioning elicits complex behaviors associated with adaptability, self-monitoring, goal-directed behavior, planning and organizing, problem-solving, decision making, impulse control, and regulating emotions (Chaytor, Schmitter-Edgecombe, & Burr, 2006; Spooner & Pachana, 2006). The results obtained from executive functioning measures provide clinical information regarding the presence or absence of neurological diseases, developmental disabilities, psychiatric conditions, or conduct disorders (Dawson, 2013; Harvey, 2012; Silver et al., 2006). The results also inform treatment response and functioning over time, functional potential and functional recovery.
educational planning, and treatment recommendations (Dawson, 2013; Harvey, 2012; Silver et al., 2006).

**Executive Functioning in Children and Adolescents**

Research examining executive functioning in children and adolescents did not emerge until the late 1980s, and even then, most of the research in this area was flawed because it was grounded in neuropsychological models, methods, and theoretical frameworks developed to evaluate executive functions in adults (Bernstein, 2009; Williams & Boll, 1997). Over the past few decades, interest in the neuropsychological assessment of children has considerably flourished due (a) to several advances in research examining neurodevelopment, the neuroanatomical organization of the cerebral cortex, and the developmental trajectory of executive functioning in preschool and school-aged children as well as (b) to increasing evidence supporting its clinical utility across settings and pediatric populations (Brocki & Bohlin, 2004; Carlson, 2003; Dawson, 2013; Garon, Bryson, & Smith, 2006; Lee, Bull, & Ho, 2013; Shing et al., 2010; Xu et al., 2013). Current literature indicates that the executive function system develops sequentially and becomes increasingly more specific with age. These findings are commensurate with other research findings that indicate the prefrontal cortex is the last cortical region of the brain to fully mature; notably, complex processes such as inhibition, working memory, and attention shifting do not appear to solidify until early adulthood (Chase et al., 2008; Jurado & Rosselli, 2007; Tsuchida & Fellows, 2009). Furthermore, evidence suggests that executive functioning represents a unitary construct in early development and a multidimensional construct in later development, which reflects the trajectory of differentiation across specific executive skills (Brydges, Reid, Fox, & Anderson, 2012; Carlson, 2003; Garon et al., 2006; Garrett, 1946; Lee et al., 2013; Shing et al., 2010).
Given that the executive function system continues to solidify across childhood and adolescence, adolescents remain at risk for organic and environmental influences that disrupt brain-behavior pathways associated with executive functioning (Harvey, 2012; Toplak et al., 2013). Current literature suggests that adolescents with compromised executive function are more likely to experience impairment across cognitive, adaptive, behavioral, and social domains as well as adverse short- and long-term outcomes in academic achievement, school readiness, social functioning and competence, and behavior (Bull & Ho, 2013; Chana, Shumb, Touloupolouc, & Chen, 2008; Chaytor et al., 2006; Muir-Broaddusa et al., 2002). Particularly, deficits in executive function compromise academic performance due to difficulties in attention, initiation, planning and organizing, and self-monitoring (Best & Miller, 2010; Carlson, 2003; Dawson, 2013; Miller, 2009). In addition, executive dysfunction is associated with poorer adaptive and coping behaviors within stress-inducing contexts, especially within the context of adverse childhood events, including poverty, toxic stress, abuse and neglect, and chronic illness (Best & Miller, 2010; Carlson, 2003; Castellanos et al., 2016; Dawson, 2013; Miller, 2009). Furthermore, executive functioning deficits are associated with difficulties in effectively processing situational information in order to successfully adapt across settings to shared rules and social roles (Barrasso-Catanzaro & Eslinger, 2016).

**The Ambiguity of Executive Functioning**

Despite the plethora of studies examining the construct of executive functioning in children and adolescents, the conceptualization and definition of the construct continue to remain elusive and inconclusive in the literature. More than 40 years ago, Luria (1973) proposed a multidimensional theory of executive functioning, which states the prefrontal cortex operates within an interconnected cortical system containing multiple sub-regions that synchronously
yield complex cognitive processes (e.g., executive functions; Luria, 1973; Semrud-Clikeman & Ellison, 2009). Evidence from a prominent study by Miyake and colleagues (2000) supported Luria’s theory and indicated that activation of executive functions requires multiple independent cortical processes to work together within an increasingly complex and interconnected functional system (Lehto, Juujarvi, Kooistra, & Pulkkinen, 2003).

Although the model supported by Luria (1973) and Miyake and colleagues (2000) remains popular within contemporary neuropsychological assessment, competing models of executive functioning have also become prominent in examining brain-behavior relationships in children and adolescents (Jurado & Rosselli, 2007; Lehto et al., 2003; Semrud-Clikeman & Ellison, 2009). A model by Baddeley and Hitch (1974) proposed that working memory processes and attentional control yield complex cognitive processes in the prefrontal cortex (Jurado & Rosselli, 2007). Norman and Shallice (1986) proposed the information processing model, a two-level hierarchical model that separates automatic simplistic cognitive processes from complex controlled cognitive processes (Jurado & Rosselli, 2007). Stuss (1992) expanded Norman and Shallice’s (1986) model and introduced a three-level hierarchical model, in which the frontal lobes facilitate increasingly complex cognitive processes via subcortical systems, complex executive functions, and self-regulation (Jurado & Rosselli, 2007). More recently, Barkley (1997) introduced a top-down model, which proposes that executive functions are involved in regulating behavior (e.g., emotional, social, etc.) within the immediate environment. Barkley hypothesized that cognitive processes involved in inhibitory behavior also regulate other executive functions such as self-regulation and goal-directed behavior (Barkley, 1997). Although the various models conceptualize the construct of executive function as a unique and complex
cognitive process, they lack clarity regarding whether the construct represents one single underlying ability, or whether it represents separate, yet unique, interrelated abilities.

**Measuring Executive Functioning in Children and Adolescents**

Given that researchers have noted the ambiguity in the construct of executive functioning in children and adolescents, especially considering the various theoretical frameworks that describe the construct, it remains challenging to develop measurement instruments that produce an accurate estimate of the targeted construct, without the undermining influence of other cognitive processes (Gioia, Isquith, Retzlaff, & Espy, 2002; Isquith, Roth, & Gioia, 2013; Miyake et al., 2000). The undermining influence is often referred to as “task impurity” or construct-irrelevant variance (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; Isquith et al., 2013; Lehto et al., 2003). Construct-irrelevant variance occurs when extraneous, uncontrollable factors compromise the validity and accuracy of test results and adversely affect the conclusions drawn from them regarding diagnosis, areas of deficit, and treatment recommendations.

**Performance-based tests.** Traditionally, performance-based tests of executive function were considered the “gold standard” for neuropsychological assessment and have been the primary modality for measuring executive function in children and adolescents (Gioia et al., 2002; Isquith et al., 2013). Performance-based tests are administered by an examiner in a highly-structured setting and utilize standardized procedures and normative methods to assess specific executive functions via the examinee’s accuracy and response time on tasks (Gioia et al., 2002; Harvey, 2012; Isquith et al., 2013). However, numerous studies indicate that several of the widely-used performance-based tests of executive function in children and adolescents, such as
the Developmental Neuropsychological Assessment- Second Edition (NEPSY-II; Korkman, Kirk, & Kemp, 2007), the Delis–Kaplan Executive Function System (D-KEFS; Delis, Kaplan, & Kramer, 2001), and the Cognitive Assessment System (CAS; Naglieri & Das, 1997), produce results that reflect weak ecological validity, which refers to the degree that an assessment instrument accurately represents behaviors in everyday environmental contexts (Spooner & Pachana, 2006).

The fact that performance-based tests consistently demonstrate weak ecological validity suggests that they do not well represent behaviors associated with the application of executive functions in real-world settings (Chaytor & Schmitter-Edgecombe, 2003; Chaytor et al., 2006). Real-world, everyday contexts generally introduce higher environmental demands, which require the independent initiation of multiple executive functions such as processing and organizing incoming stimuli and planning responses (Castellanos et al., 2016; Gioia et al., 2002; Isquith et al., 2013; Toplak et al., 2013). In contrast, the standardized procedures of performance-based tests employ systematic prompts and cues to alter the examinee’s problem-solving approach; such prompts and cues arguably interfere with the examinee’s ability to independently initiate and fully engage in executive functions and spoil the novelty of tasks following the first exposure (Chana et al., 2008; Gioia et al., 2002; Isquith et al., 2013; Toplak et al., 2013). Moreover, results from performance-based tests lack generalizability to behavior exhibited in real-world contexts because the tests were developed to measure specific executive functions and not necessarily the multidimensional, interconnected nature of executive functions (Chana et al., 2008; Isquith et al., 2013; Toplak et al., 2013).

**Behavior rating scales.** The introduction of behavior rating scales targeting executive function in the late 1990s revolutionized the approach to measuring the construct and increased
confidence in obtaining ecologically valid results from neuropsychological evaluations (Gioia et al., 2002; Castellanos et al., 2016; Toplak et al., 2013). Since the implementation of behavior rating scales in research and practice, studies have indicated that these scales demonstrate stronger ecological validity than performance-based tests of executive function (Chaytor et al., 2006; Gioia et al., 2002; Spooner & Pachana, 2006; Toplak et al., 2013). Additional strengths of behavior rating scales include their ability to unobtrusively measure behaviors; to measure the frequency of targeted behaviors within a specific time frame, as opposed to one testing session; to sample multiple sources (i.e., parent, teacher, and self) for behavioral information; and to more accurately assess functional outcomes across contexts (Chaytor et al., 2006; Gioia et al., 2002; Spooner & Pachana, 2006; Toplak et al., 2013). Given the noted strengths of behavior rating scales, converging evidence suggests that such scales, utilized within a multimethod, multisource approach to assessment, aid in accurately predicting a child’s level of everyday executive function within and across multiple contexts and functional domains (Castellos et al., 2016; Chaytor et al., 2006; Gioia et al., 2002; Spooner & Pachana, 2006).

Although several behavior rating scales have been developed to measure executive functioning in children and adolescents, the Behavior Rating Inventory of Executive Function (BRIEF; Gioia, Isquith, Guy, & Kenworthy, 2000) is recognized as the first rating scale developed to assess everyday executive functioning in children and adolescents (Castellos et al., 2016; Gioia, Isquith, & Guy, 2001; Gioia et al., 2002; Spooner & Pachana, 2006). Given the clinical utility and increasing popularity of the BRIEF rating scales, the Behavior Rating Inventory of Executive Function–Self-Report Form (BRIEF-SR; Guy, Isquith, & Gioia, 2004) was developed for adolescents ages 11 to 18 years in order to supplement parent and teacher ratings on the BRIEF. Items across the BRIEF rating scales targeted eight specific executive
functions: Inhibit, Shift, Emotional Control, Working Memory, Plan/Organize, Organization of Materials, Monitor, and Initiate or Task Completion. In addition, the BRIEF rating scales yield two broad indices, the Behavioral Regulation Index and the Metacognition Index, as well as one total score, the Global Executive Composite.

A revision of the BRIEF (Gioia et al., 2000), the Behavior Rating Inventory of Executive Function, Second Edition (BRIEF2; Gioia, Isquith, Guy, & Kenworthy, 2015a) was published in 2015. Three BRIEF2 forms were developed, including the parent and teacher forms completed for adolescents ages 5 to 18 years and the self-report form completed by adolescents ages 11 to 18 years. The aim of the revision was to maintain the theoretical framework of the original scale while improving its administration and scoring procedures, increasing its clinical and research utility and its accuracy and specificity in identifying executive function deficits, and updating its internal structure according to current research (Gioia, Isquith, Guy, & Kenworthy, 2015b). The BRIEF2 has a similar score structure to its predecessor.

Relatively recently, the Comprehensive Executive Function Inventory (CEFI; Naglieri & Goldstein, 2013a) was published to assess executive functioning in children and adolescents ages 5 to 18 years. Three CEFI forms were developed, including the parent and teacher forms completed for adolescents ages 5 to 18 years and the self-report form completed by adolescents ages 12 to 18 years (Climie, Cadogan, & Goukon, 2014; Fenwick & McCrimmon, 2015; Naglieri & Goldstein, 2013b). The items across the CEFI rating scales target nine specific executive functions: Attention, Inhibitory Control, Flexibility, Emotion Regulation, Planning, Organization, Working Memory, Self-Monitoring, and Initiation. Notably, the CEFI maintains a parallel score structure across the parent, teacher, and self-report forms.
Psychometric Standards and Behavior Rating Scales Measuring Executive Functions

According to The Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014), the psychometric properties of an assessment instrument are paramount as these properties affect how resulting scores influence decisions made in practice and research. One such psychometric property, reliability, significantly increases the confidence that can be placed in such scores. Reliability refers to the consistency and dependability of scores from the same measurements across replications of a testing procedure (AERA, APA, & NCME, 2014, p. 33). In contrast to reliability, measurement error refers to the influence of confounds introduced in the representation of a targeted construct. Thus, measurement error undermines trust in scores obtained from an assessment instrument.

Classical Test Theory (CTT) is considered the traditional model for understanding reliability and measurement error in educational and psychological measures (Briesch, Chafouleas, & Johnson, 2016). CTT assumes that any obtained score reflects two components: (1) a theoretical true score, which represents true differences across individuals on a measured construct and (2) measurement error, which represents the unavoidable influence of possible confounds, ranging from random error to systematic influences. As such, methods based on CTT allow for evaluations to determine whether a measurement tends to be stable (e.g., across items, time, or rater) for a particular behavior as well as the extent of measurement error affecting obtained scores (Briesch et al., 2016; Prion, Gilbert, & Haerling, 2016). Reliability analyses based on CTT produce indices reflecting the amount of observed score variance that is attributable to true score variance. These reliability analyses have primarily targeted only one source of systematic measurement error at a time.
Three types of reliability analyses are widely used to assess the reliability of not only social, emotional, and behavioral rating scales, but also rating scales of executive function. Internal consistency analyses measure the degree to which instrument items proposed to measure the same construct produce similar scores and evaluate error variance due to item-specific factors. For example, poor item quality (e.g., length and linguistic complexity) may affect examinee responses, and inadequate item content sampling likely results in an inaccurate representation of the proposed construct. Test-retest reliability analyses measure the consistency across test scores obtained over time and evaluate error variance associated with inaccuracy in measurement at each time point. Factors such as the differences in the examinee’s internal states (e.g., mood, motivation, and fatigue) across time and inconsistency in the construct being measured (e.g., feelings of anxiety) may influence the variability in scores over time. Inter-rater reliability analyses measure consistency across raters and evaluate error variance due to differences attributable to the raters. Internal and external factors such as rater biases and response patterns, scoring errors, and ambiguous scoring procedures may lead to variability in raters’ scores. When such analyses are not carefully controlled, differences due to variation in the experience of raters who are observing behaviors in different settings may be falsely attributable to the raters, per se. In addition to these sources of error traditionally evaluated using CTT, several other sources of error, such as source and setting, instrument, and dimension, are known to affect informant ratings of executive function (De Los Reyes & Kazdin, 2005; Isquith et al., 2013; Russell, Russell, & Hill, 2005; Spinella, 2005). Measurement error attributable to source and setting, instrument, and dimension are described in the sections that follow.

**Source and setting variance.** When considering informant ratings of executive function, it is important to consider the types of error variance evaluated during inter-rater reliability
analyses. Source variance refers to the measurement error attributable to informant response bias, as “source” refers to the source of information. Informant response bias generally accounts for a significant amount of error in scores from behavior rating scales—especially when informants rate behaviors they have observed in the past—as responses require the application of memory and inference (Lakes & Hoyt, 2009). Three types of informant response biases related to behavior rating scales include (a) rating the behavior in an overly positive or negative manner (i.e., the halo effect), (b) rating the behavior in extremes, yielding consistently higher or lower scores than warranted (i.e., the severity effect), or (c) consistently rating the behavior as neutral (i.e., the central tendency effect; Merrell, 2003).

Setting variance refers to the measurement error attributable to the discrepancy in ratings across different settings in which the informants observe behaviors; it is associated with the situational specificity of behavior, which refers to adolescents exhibiting different patterns of behavior in different settings (e.g., home and school; Karver, 2006). In most cases of use of behavior rating scales, source and setting variance are inextricably entangled; parents observe behavior in the home and community, and teachers observe behavior in the school setting. Although employing multiple caregivers observing behavior in the same setting (e.g., two teachers in the same classroom) may allow for disentangling setting effects from source effects, self-ratings always reflect the experience of one source across settings. Such entanglement of source and setting influences is evident in the literature reviewed in the paragraphs that follow.

A widely-used method for measuring the discrepancy between different raters in rating scale research is an inter-rater reliability analysis. As noted previously in the section focused on CTT, inter-rater reliability coefficients estimate relative consistency in measurement and its inverse, error attributable to rater bias and random response error. Studies that have examined the
inter-rater reliability of multidimensional social, emotional, adaptive, and behavioral rating scales indicate that reports of children’s and adolescent’s behavior between pairs of raters who serve in similar roles and settings (e.g., pairs of teachers and pairs of caregivers) generally yield moderately high mean correlations ($r \sim .60$; Achenbach, 2011; De Los Reyes & Kazdin, 2005; Duhig, Renk, Epstein, & Phares, 2000; Grietens et al., 2003). However, several studies have consistently indicated that the reports of children’s and adolescent’s behaviors across different raters from different settings (e.g., between parent and teachers, between teachers and children, between parents and children) yield lower correlations ($r \sim .40$) than the moderately low correlations obtained from the reports from pairs of raters (Achenbach, 2011; De Los Reyes & Kazdin, 2005; Duhig et al., 2000; Grietens et al., 2003; Youngstrom, Loeber, & Stouthamer-Loeber, 2000). The consensus currently accepted among researchers and clinicians suggests that inconsistencies between multiple raters’ reports are likely attributed to situational specificity and to raters’ interpersonal relationships with those being rated (e.g., child or student; Achenbach, 2011; De Los Reyes & Kazdin, 2005; Duhig et al., 2000; Grietens et al., 2003). A classic study in rating scale research conducted by Achenbach, McConaughy, and Howell (1987) employed a meta-analysis of multiple raters’ reports on various rating scales for children and adolescents. The study’s findings indicated that a pattern of low correspondence between raters’ reports of adolescents’ behaviors is particularly evident for parent-child and parent-adolescent raters (Achenbach et al., 1987).

Although evidence generally supports low parent-adolescent correspondence on behavior rating scales, findings have shown relatively higher correspondence in parent-adolescent reports of salient behavior constructs as opposed to their reports of less salient behavior constructs. Salient behaviors are objective and readily observable such as externalizing behavior, including
aggression, delinquency, and hyperactivity, and adaptive behavior, including independent daily behaviors (Karver, 2006). Less salient behaviors are subjective and not directly observed such as internalizing behavior, including sadness, withdrawal, and anxiety, and social behavior, including likeability and peer perceptions (Achenbach et al., 1987; Achenbach, 2011; Barker, Bornstein, Putnick, Hendricks, & Suwalsky, 2007; De Los Reyes & Kazdin, 2005; Duhig et al., 2000; Grietens et al., 2003; Karver, 2006; Youngstrom et al., 2000). Several studies have supported relatively higher correspondence between parent-adolescent reports of externalizing and adaptive behavior than of internalizing and social behavior, though the correspondence still remains low ($r \sim .30$; Duhig et al., 2000; Grietens et al., 2003; Karver, 2006; Youngstrom et al., 2000). A pattern of consistently low to moderate correlations ($r \sim .20$ to $.30$) has also been shown between parent-adolescent reports of social behavior (Gresham, Elliott, Cook, Vance, & Kettler, 2010; Renk & Phares, 2004). Researchers and clinicians attribute lower cross-informant agreement for internalizing and social constructs because the constructs require raters to subjectively decide whether or not these less salient behaviors exist, which inevitably presents barriers to assessing the frequency, severity, and duration of the adolescent’s behavior, and its impact on the adolescent’s overall functioning (De Los Reyes & Kazdin, 2005; Youngstrom et al., 2000). Furthermore, researchers suggest that it may be more difficult for parents to attend to everyday internalizing and social behaviors because they are much less consequential within the home environment than externalizing or maladaptive behaviors (Renk & Phares, 2004).

Although the literature in rating scale research for children’s and adolescent’s social, emotional, adaptive, and behavioral functioning has flourished since the 1990s, relatively little research exists that has examined agreement in cross-informant ratings of executive function in adolescents and in particular, the concordance and discordance across parent-adolescent reports.
of executive function. Consistent with previous findings in rating scale research, converging evidence has also indicated a pattern of low parent-adolescent correspondence in ratings of executive function (Baron, 2000; Walker & D’Amato, 2006). Specifically, the parallel subscales across the BRIEF (Gioia et al., 2000; Gioia et al., 2004) Parent and Self-Report forms evidenced low to moderately high correlations (e.g., \( r \approx .36 \) to .57; Baron, 2000; Walker & D’Amato, 2006). Furthermore, moderately high correlations between parent-adolescent reports were found for both of the BRIEF indices, the Behavior Regulation Index \( (r = .52) \) and the Metacognition Index \( (r = .57) \), and the total score, the Global Executive Composite \( (r = .56; \) Guy et al., 2004).

The BRIEF2 (Gioia et al., 2015a) evidenced a similar pattern of low to moderately high correlations between parent-adolescent reports in clinical and nonclinical samples (Gioia et al., 2015b). For the BRIEF2 nonclinical sample, the Self-Monitor, Shift, and Emotional Control subscales yielded moderately high correlations (i.e., ranging from .46 to .59) across raters, whereas the Inhibit, Working Memory, and Plan/Organize subscales yielded higher correlations (i.e., ranging from .62 to .67; Gioia et al., 2015b). Furthermore, all three of the BRIEF2 indices, Behavior Regulation, Cognitive Regulation, and Emotion Regulation, as well as the total score, Global Executive Composite, yielded moderately high to high correlations between parent-adolescent reports from a nonclinical sample (i.e., ranging from .59 to .74; Gioia et al., 2015b). In comparison, the BRIEF2 clinical sample evidenced lower parent-adolescent correlations (i.e., ranging from .20 to .35) across six subscales, including Self-Monitor, Shift, Emotional Control, Inhibit, Working Memory, and Plan/Organize. Additionally, lower correlations (i.e., ranging from .25 to .35) were also found for the Behavior Regulation, Cognitive Regulation, and Emotion Regulation indices and the Global Executive Composite. Similarly, findings have
shown that the CEFI (Naglieri & Goldstein 2013a) evidences low consistency across parent-adolescent raters (i.e., correlations ranging from .38 to .67; Naglieri & Goldstein 2013b).

Similar to the difficulties associated with observing internalizing symptoms and social behavior, it is difficult for parents to directly observe and interpret cognitive processes associated with executive functions as well as functional outcomes associated with executive function deficits. Particularly, behaviors associated with executive function deficits, such as poor attention, memory, emotion regulation, and self-monitoring, are less salient and less overtly bothersome than externalizing behaviors, and thus, often go unnoticed by parents until significant and apparent declines occur in other functional domains (e.g., home, school, community, and social; Isquith et al., 2013). Additionally, due to the fact that executive functions are not directly observable, parents are likely to misinterpret behavioral manifestations of these cognitive processes due to multiple internal and external factors that affect parents’ perspectives of the behavior. For example, De Los Reyes and Kazdin (2005) noted that parental factors such as (a) the frequency of behavior; (b) the type of behavior problem; (c) the severity of the impairment; (d) the context in which the behavior is observed; (e) the parents’ attributions for specific behaviors; and (f) and the parents’ internal states (e.g., psychopathology, stress, and fatigue) and external influences (e.g., immediate environment) significantly alter reports of presenting behavior. In addition, children and adolescents with moderate to significant deficits in executive function typically lack awareness of their own cognitive deficits and of how these deficits negatively impact functioning across home, school, and social domains, which inevitably contributes to discrepancies between parent-adolescent reports of everyday executive function (Best & Miller, 2010; Willard et al., 2016). Furthermore, children’s and adolescent’s developmental and cognitive level present barriers for translating their level of awareness into
accurate and valid self-reports of executive functioning (Willard et al., 2016). Given that the literature suggests that executive functioning and self-awareness increase with age, it is plausible that older adolescents have increased awareness of their executive function deficits and therefore, endorse more difficulties on self-report scales than younger adolescents (Best & Miller, 2010).

**Instrument variance.** Instrument variance refers to the inconsistency between test scores from different rating scales proposed to measure the same construct and administered simultaneously. Instrument variance is related to differences in the items and general methods with which different instruments assess a construct. Thus, systematic measurement error attributable to instrument variance may be due to differences in item content or wording, inconsistency in the rating scale’s response format, or differences in assessment method (self-report versus direct observation; Merrell, 2003). Consistency in test scores across multiple instruments proposed to measure the same construct provides evidence for convergent validity. Convergent validity, a subtype of criterion-related validity, refers to the degree to which informants’ ratings are similar on instruments proposed to measure theoretically similar constructs. As such, instruments proposed to measure the same underlying theoretical construct should yield generally consistent results, with moderate to high inter-instrument correlations (Greenbaum, Dedrick, Prange, & Friedman, 1994).

According to the research examining the convergent validity of parent rating scales, there is a pattern of moderate concordance between scores from rating scales of executive function and scores from widely used rating scales targeting related constructs. This literature is quite sparse and almost solely focused on the relations between executive function and behaviors associated with attention-deficit/hyperactivity disorder (ADHD; Gioia et al., 2000; Goulden & Silver, 2009;
Jarratt, Riccio, & Siekierski, 2005). For example, convergent validity studies have revealed moderate correlations between scores from the BRIEF (Gioia et al., 2000) Parent form and (a) the Inattention and Hyperactivity/Impulsivity scales from the ADHD-Rating Scale- Fourth Edition (ADHD-RS-IV; DuPaul et al., 1998); (b) the Attention Problems scale from the Child Behavior Checklist (CBCL; Achenbach, 1991); (c) the Inattention and Hyperactivity scales from the Behavior Assessment System for Children (Reynolds & Kamphaus, 1992); and (d) the relevant scales from the Conners Rating Scale (Conners, 1989).

Similarly, convergent validity studies also provide evidence of moderate correspondence between scores from the BRIEF2 (Gioia et al., 2015a) Parent form and behavior scales associated with ADHD from other rating scales. Generally moderate correlations (i.e., ranging from .24 to .76) have been shown between the BRIEF2 parent subscales and the ADHD-RS–IV Inattention and Hyperactivity scales (DuPaul et al., 1998). Additionally, the correlations between the BRIEF2 parent subscales and the CBCL (Achenbach, 1991) Attention Problems and Rule-Breaking scales were weak (i.e., ranging from .14 to .34). However, the correlations between the BRIEF2 parent subscales and the BASC-2 (Reynolds & Kamphaus, 2004) Hyperactivity and Inattention scales were slightly stronger (i.e., ranging from .25 to .81). Similarly, findings also have shown weak to moderate correlations between the BRIEF2 parent subscales and the Inattention and Hyperactivity/Impulsivity scales from the Conners 3 Rating Scale (Conners 3; Conners, 2008) (i.e., ranging from .10 to .74).

In comparison, convergent validity studies provide evidence of generally weak to moderate correspondence between scores from the BRIEF2 (Gioia et al., 2015a) Parent form and relevant executive function scales from other rating scales; however, very few studies target the actual construct of executive function. For example, findings have shown weak correlations
between the BRIEF2 parent subscales and the CBCL (Achenbach, 1991) Thought Problems scale (i.e., ranging from .18 to .29). One study reported weak to moderate correspondence between the BRIEF2 parent subscales and the Executive Function scale from the Conners 3 (Conners, 2008). Specifically, moderate correlations were found between the BRIEF2 Initiate, Working Memory, Plan/Organize, and Organization of Materials subscales (i.e., ranging from .54 to .70), whereas much lower correlations were found between the BRIEF2 Inhibit, Self-Monitor, Shift, Emotional Control, and Task-Monitor subscales (i.e., ranging from .17 to .37; Gioia et al., 2015b). Furthermore, only one convergent validity study to date has provided evidence of moderate correspondence between the total scores from the BRIEF and the CEFI (Naglieri & Goldstein, 2013a) parent rating scales. In a study conducted with parents from a clinical subsample, moderate correlations ($r = .76$) were found between the total scores obtained from parents’ reports on the BRIEF and CEFI rating scales (Naglieri & Goldstein, 2013b).

The literature focusing on the convergent validity of parent rating scales is quite sparse, but the literature focusing on the same evidence across self-report rating scales is even less developed. Only two studies providing evidence of weak to moderate relations between executive function and ADHD behavior scales were presented in the BRIEF2 Professional Manual (Gioia et al., 2015b). A pattern of low to moderate correlations was evidenced between the BRIEF2 (Gioia et al., 2015a) Self-Report subscales and the BASC-2 (Reynolds & Kamphaus, 2004) Self-Report of Personality Inattention and Hyperactivity scales (i.e., ranging from .38 to .73). However, slightly lower correlations were found between the BRIEF2 Self-Report subscales and the CBCL Youth Self-Report (CBCL-YSR; Achenbach, 1991) Attention Problems and Rule-Breaking scales (i.e., ranging from .26 to .59).
Unfortunately, the evidence provided for relations between self-reports of executive function is minimal. For example, generally weak correspondence was found between the BRIEF2 Self-Report subscales and the CBCL-YSR Thought Problems scale (i.e., correlations ranging from .30 to .53). In addition, one convergent validity study conducted with adolescents from a clinical subsample has provided evidence of moderately high correlations \((r = .79)\) between the total scores obtained from adolescents’ reports on the BRIEF and CEFI (Naglieri & Goldstein, 2013a) self-report rating scales.

**Dimension variance.** Behavior rating scales typically produce scores measuring somewhat specific dimensions as well as more global scores measuring general dimensions. This fact is certainly true for behavior rating scales measuring executive functions. For example, as previously stated, the internal structure of the BRIEF2 (Gioia et al., 2015a) and the CEFI (Naglieri & Goldstein, 2013a) is hierarchical, though the BRIEF2’s organizational structure targets specific dimensions of executive function—or subscales—that contribute to broad areas of executive function and global executive functioning, and the CEFI’s organizational structure targets subscales of executive function that only contribute to global executive functioning. From such an organization, it is often unclear exactly how much of the variation in scores across a rating scale is due to the ways the more specific dimensions are measured by subscales. If all subscales vary little for individuals, using the more global scores alone may be satisfactory.

With these issues in mind, dimension variance refers to variation across test scores targeting similar—yet not identical—behaviors. Thus, inconsistency across subscale scores may be attributable to the subscales measuring different specific dimensions (Briesch et al., 2014). This inconsistency reflects discriminant validity, a subtype of criterion-related validity, across subscales. Discriminant validity refers to the degree to which test scores proposed to measure
distinct constructs diverge in measurement. It is expected that subscales proposed to measure the same underlying theoretical construct should yield generally consistent results, with moderate to high inter-subscale correlations (Greenbaum et al., 1994), whereas subscales proposed to measure different constructs should yield generally divergent results, with low inter-subscale correlations.

Dimension variance is particularly relevant to understanding systematic variance across specific domains of executive function, as the literature suggests that there are many different indicators of executive functioning due to its broad “umbrella” term that refers to a specific group of cognitive processes that facilitate higher-order thinking and executive control. As evident in the score structures outlined earlier, the BRIEF2 (Gioia et al., 2015a) and the CEFI (Naglieri & Goldstein, 2013a) have adopted this multifaceted concept of executive function, in which specific domains of executive function serve as indicators for the broader construct of executive function.

Studies presented in the BRIEF2 Professional Manual (Gioia et al., 2015b) and the CEFI Technical Manual (Naglieri & Goldstein, 2013b) suggest that both rating scales evidence strong relations between the subscales and the total score. The intercorrelations between all subscales, indices, and the total score are based on the norming samples for the BRIEF2 (Gioia et al., 2015a) Parent and Self-Report forms. The intercorrelations presented for the BRIEF2 Parent form indicated that the nine subscales evidence moderately low to strong intercorrelations (ranging from .41 to .85). Furthermore, the subscales had moderate to strong correlations with the three indices, including the Behavior Regulation Index (ranging from .59 to .95), Emotion Regulation Index (ranging from .47 to .94), and the Cognitive Regulation Index (ranging from
Moderate to strong correlations (ranging from .75 to .95) were also found between the subscales and the total score, the Global Executive Composite.

Similarly, the intercorrelations presented for the eight subscales from the BRIEF2 (Gioia et al., 2015a) Self-Report form indicated moderate to strong relations between subscales, indices and the total score, with slightly higher intercorrelations compared to the BRIEF2 Parent form (Gioia et al., 2015b). Specifically, the Self-Report form evidenced moderate to strong correlations between all subscales (ranging from .52 to .94) and between the subscales and the indices, including Behavior Regulation (ranging from .70 to .94), Emotion Regulation (ranging from .70 to .92), and Cognitive Regulation (ranging from .62 to .94). Additionally, the intercorrelation matrix indicated moderate to strong relations between the subscales and the Global Executive Composite (correlations ranging from .78 to .95). These results indicate that the BRIEF2 subscales consistently measure a global construct of executive function, which suggests that the total score, rather than the subscale scores, may be a better indicator of the underlying theoretical construct of executive function.

Confirmatory factor analysis (CFA) directly addresses variance due to specific dimensions. Although it does so at the latent-variable level where all variance is reliable, true-score variance, this method informs conclusions about dimension variance in the use of rating scale data. For example, CFAs based on the norming samples for the BRIEF2 (Gioia et al., 2015a) Parent and Self-Report forms provide further evidence of little unique variance attributable to subscale scores and thus, less multidimensionality in the construct of executive function (Gioia et al., 2015b). According to the CFA for the BRIEF2 Parent form, a three-factor solution best fit the data. The results from the CFA, which employed all nine subscales as measured variables, indicated strong relations (correlations ranging from .82 to .92) between the
three factors (e.g., Behavior Regulation, Emotion Regulation, and Cognitive Regulation). These results suggest that the subscales are measuring the same general dimension and only somewhat measuring specific dimensions. Furthermore, results indicated strong factor loadings for the Behavior Regulation factor and the Inhibit (.87) and Self-Monitor (.83) subscales. A similar pattern of strong relations was found between the Emotion Regulation factor and the Shift (.85) and Emotional Control (.83) subscales. Additionally, strong relations were found for the Cognitive Regulation factor and the Initiate (.89), Working Memory (.89), Plan/Organize (.92) and Organization of Materials (.84) subscales, and a slightly weaker relation was found for the Task-Monitor (.79) subscale.

The CFA for the BRIEF2 (Gioia et al., 2015a) Self-Report form indicated that a three-factor solution best fit the data. The CFA, which employed all seven subscales as measured variables, revealed strong relations (correlations ranging from .89 to .96) between the Behavior Regulation, Emotion Regulation, and Cognitive Regulation factors, again suggesting they are measuring the same general dimension and only somewhat measuring specific dimensions. A strong factor loading was found between the Behavior Regulation factor and the Inhibit (.86) subscale, and a slightly weaker factor loading was found for the Self-Monitor (.79) subscale. Similarly, strong relations were found between the Emotion Regulation factor and the Shift (.85) subscale and a slightly weaker relation was found for the Emotional Control (.77) subscale. The Cognitive Regulation factor evidenced strong relations with the Task Completion (.88), Plan/Organize (.91), and Working Memory (.89) subscales.

In contrast to the BRIEF2 (Gioia et al., 2015a) analyses using CFA, scale-level exploratory factor analysis (EFA) for the CEFI (Naglieri & Goldstein, 2013b) Parent and Self-Report forms provides evidence of strong relations between the nine subscales and a single
general factor for each scale. The scale-level EFA for the CEFI Parent form revealed that all nine subscales loaded strongly on the factor (with factor loadings ranging from .84 to .96). Similarly, the scale-level EFA for the Self-Report form revealed that all nine subscales evidenced moderate to strong factor loadings on the factor (with factor loadings ranging from .72 to .89). Like the correlations on which these structural analyses were based, these findings indicate that little unique variability exists in BRIEF2 and CEFI subscale scores. The construct of executive function likely reflects a unidimensional construct—rather than a multidimensional construct—thus, providing evidence that a global score is a better indicator of the construct of executive function.

**Purpose of the Study**

The purpose of the current study is to expand the literature examining reliability, validity, and generalizability of test scores obtained from the BRIEF2 (Gioia et al., 2015a) and the CEFI (Naglieri & Goldstein, 2013a) rating scales of executive function in adolescents. In particular, this research will provide further insight into the sources of variance that disrupt the dependability of scores from rating scales of executive function. Currently, CTT has been the only model utilized to examine the sources of error variance from these rating scales. A significant limitation in employing CTT is that each analysis (e.g., inter-rater reliability) typically allows for examination of only one source of error variance at a time and only yields a global error term, which assumes that only one source of measurement error accounts for variance in test scores (Briesch et al., 2014). This assumption of CTT is undoubtedly false, as multiple sources of measurement error account for variance in test scores, particularly when assessment administration includes multisource and multimethod approaches.
Generalizability theory (GT; Cronbach, Rajaratnam, & Gleser, 1963) expands CTT methods by enabling researchers to examine multiple sources of error simultaneously via the estimation of variance components (Brennan, 2011; Briesch et al., 2016). GT analyses aid researchers in distinguishing between random error, which refers to unpredictable fluctuations in scores due to unexplainable and uncontrollable confounds, and systematic error, which refers to consistent fluctuations in scores due to construct-irrelevant sources. The estimation of variance components provides information regarding how multiple sources of variance influence scores as well as how these sources of variance interact (Brennan, 2011; Briesch et al., 2016). GT also enables the calculation of generalizability coefficients, which expands the study’s implications and applicability. Generalizability coefficients indicate how accurately the study’s findings generalize from the object of measurement (e.g., observed behavior, or in this case, general executive functioning) to the universe of admissible observations (e.g., the range of potential conditions to measure a targeted construct; Brennan, 2011; Briesch et al., 2016).

The current study employed GT analyses to compare and contrast the sources of error contributing to test score variance in the BRIEF2 (Gioia et al., 2015a) and the CEFI (Naglieri & Goldstein, 2013a) rating scales among adolescents and their parents. The study examined the possible sources of variance due to systematic error variance attributable to the raters, the instruments, the instruments’ targeted dimensions, and their interactions (Brennan, 2011; Briesch et al., 2014). The study enhanced the existing literature regarding the construct validity of executive function rating scales, as it (a) investigated the construct validity of specific executive function domains and (b) compared and contrasted the proportion of variance attributable to specific executive function domains as opposed to the overall executive function domain.
In addition to identifying the sources of error that contribute to test score variance in the BRIEF2 (Gioia et al., 2015a) and the CEFI (Naglieri & Goldstein, 2013a) rating scales, this study further expanded the relatively small literature that exists in adolescents’ self-reports of executive function. According to the literature, it appears that studies have found mixed results with regard to the validity and reliability of adolescent self-reports of their own social, emotional, and behavioral functioning. On one hand, findings have indicated that children ages 6 to 11 years are generally better reporters than adolescents ages 12 to 19 years because adolescents tend to either under- or over-report symptoms and problems due to factors associated with social desirability or heightened saliency in the perception of their own behaviors, respectively (De Los Reyes & Kazdin, 2005; Grietens, 2004; Kramer, 2006). On the other hand, studies have indicated that adolescents’ self-reports actually become increasingly valid, insightful and more useful than parent or teacher ratings because the quality of adolescents’ interpersonal relationships with adults shift such that adolescents generally prefer to spend more time with their peer group than with adults and may not openly share the severity or frequency of behavioral, social, and emotional problems with others, and especially with adults (Baker et al., 2007; Karver, 2006; Willard et al., 2016; Youngstrom et al., 2003). Given these discrepant findings as well as the necessity of behavior rating scales in comprehensive assessment, and especially in neuropsychological assessment, it becomes increasingly valuable to evaluate the role of behavior rating scales in adolescents’ self-reports of executive function.

The primary research question of this study aims to answer what amount of error variance is attributed to the instrument, the rater, and the instrument-by-rater interaction in parent-adolescent reports of executive function. The secondary research question further expands the study and aims to answer what amount of error variance is attributed to dimension, the
instrument-by-dimension interaction, the rater-by-dimension interaction, and the rater-by-
instrument-by-dimension interaction beyond the error variance due to the instrument, the rater,
and the instrument-by-rater interaction in parent-adolescent reports of executive function. Based
on previous rating scale research (Bergeron et al, 2008; De Los Reyes & Kazdin, 2005; Duhig, et
al., 2000; Grietens et al., 2003; Kramer, 2006), it is hypothesized that the rater component will
contribute the largest proportion of variance in executive functioning. Consistent with previous
research in executive function (Reynolds et al., 2009), it is also hypothesized that the dimension
component will contribute to a large proportion of variance in executive functioning, though not
as large as the proportion attributable to the rater component. Furthermore, it is hypothesized that
the instrument component will contribute a small proportion of variance in executive
functioning. Lastly, it is hypothesized that the dependability coefficient will be less than .80,
which is the typical lower level boundary for acceptable reliability.

**Method**

**Participants**

Participants included 42 adolescents ages 12 to 17 and one of their caregivers recruited from
the population of students attending a public middle and high school located in a rural school
district and enrolled in sixth- through twelfth-grade in general education or special education.
Participant recruitment occurred during the winter months (mid-January to early-March) of 2018
and targeted parent-adolescent dyads of adolescents enrolled in selected public schools.

The sample included 42 adolescents between the ages of 12 to 17 years ($M = 14.5, SD =
1.9$ years). Of this sample, 81% of adolescents were female ($n = 34$). Twenty-nine percent of
adolescents were enrolled in 6th grade, 7% in the 7th grade, 21% in the 8th grade, 19% in the 9th
grade, 14% in the 10th grade, 2% in the 11th grade, and 7% in the 12th grade. In terms of
race/ethnicity, the majority of adolescent participants were identified as White (62%). Twenty-nine percent identified as Black, 12% identified as otherwise classified (Latina or Latino, biracial, Native American, and multiracial), and 2% identified as Asian/Pacific Islander. Five percent of adolescents were of Hispanic origin and 100% of adolescents spoke English as their primary language. Of the sample, 14% \((n = 6)\) of adolescents received special education services under the classifications of Gifted \((n = 2)\), Other Health Impaired \((n = 1)\), and Speech Impaired and/or Language Impaired \((n = 2)\). Two adolescents presented with a diagnosis from the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (APA, 2013), including ADHD and an anxiety disorder.

The sample included caregivers (5 males, 37 females) between the ages of 32 to 71 years \((M = 42.7, SD = 7.5 \text{ years})\). Of the caregivers completing the forms, 79% \((n = 33)\) identified as mothers, 12% \((n = 5)\) identified as fathers, and 10% \((n = 4)\) identified as otherwise classified (e.g., grandparent or aunt). In terms of race/ethnicity, 64% identified as White, 26% identified as Black, 7% identified as otherwise classified (i.e., Latina or Latino, biracial, and multiracial), and 2% identified as Asian/Pacific Islander. Two percent of caregivers were of Hispanic origin. A total of 95% of caregivers spoke English as their primary language; Spanish \((n = 1)\) and Chinese \((n = 1)\) were reported as the other primary languages. The majority of caregivers obtained a high school degree (40%); sixteen percent obtained a specialist degree (i.e., trade school, nursing, etc.) or an associate’s degree, 14% obtained a 4-year college degree, 12% obtained a graduate degree (i.e., master’s and doctoral degree), and 9% obtained a general equivalency diploma.

**Measures**

**Behavior Rating Inventory of Executive Function, Second Edition (BRIEF2).** The BRIEF2 (Gioia, et al., 2015a) is a rating scale used to measure behaviors associated with
executive functioning in children and adolescents ages 5 to 18 years. For the purpose of this study, only the BRIEF2 Parent form and the BRIEF2 Self-Report form were employed. In addition, only the scores derived from six subscales (Inhibit, Shift, Self-Monitor, Emotional Control, Plan/Organize, and Working Memory) and the Global Executive Composite—which are common across the BRIEF2 forms—were analyzed. All scores are T Scores ($M = 50; SD = 10$) based on gender and age group. For the BRIEF2 Parent form, the norms are based on four age groups (i.e., norm blocks for ages 5 to 7, 8 to 10, 11 to 13, and 14 to 18). For the BRIEF2 Self-Report form, the norms are based on two age groups (i.e., norm blocks for ages 11 to 14 and 15 to 18). See Table 1 for a description of these scores.

Table 1

<table>
<thead>
<tr>
<th>Index or Scale</th>
<th>Number of Items Parent</th>
<th>Number of Items Self-Report</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inhibit</td>
<td>8</td>
<td>8</td>
<td>The ability to control and resist an impulse, including effectively inhibiting one’s own behavior during appropriate times.</td>
</tr>
<tr>
<td>Shift</td>
<td>8</td>
<td>8</td>
<td>The ability to transition and direct attention according to situational demands, including appropriately switching between activities, topics, or a problem-solving approach.</td>
</tr>
<tr>
<td>Self-Monitor</td>
<td>4</td>
<td>5</td>
<td>The ability to monitor the effect of own behavior on others.</td>
</tr>
<tr>
<td>Emotional Control</td>
<td>8</td>
<td>6</td>
<td>The ability to effectively and appropriately regulate emotional responses, including regulating one’s fluctuations in mood, emotional outbursts, and emotional reactions.</td>
</tr>
</tbody>
</table>
Table 1 (Continued)

<table>
<thead>
<tr>
<th>Index or Scale</th>
<th>Number of Items Parent</th>
<th>Number of Items Self-Report</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plan/Organize</td>
<td>8</td>
<td>10</td>
<td>The ability to effectively manage current and future tasks, including goal setting, developing detailed plans to complete short and long-term tasks, and summarizing verbal and nonverbal information necessary for understanding and effectively communicating main concepts and ideas.</td>
</tr>
<tr>
<td>Working Memory</td>
<td>8</td>
<td>8</td>
<td>The ability to mentally hold and manipulate information in order to sustain attention to presented tasks and to complete multistep instructions.</td>
</tr>
<tr>
<td>General Executive Composite</td>
<td>63</td>
<td>55</td>
<td>This score represents an estimate of a child’s or adolescent’s overall level of executive functioning across subscales.</td>
</tr>
</tbody>
</table>

The BRIEF2 (Gioia et al., 2015a) Parent form was developed for parents to rate problem behaviors that have occurred over the past 6 months for children ages 5 to 18 years. The BRIEF2 Parent form contains a total of 63 items that are rated on a 3-point scale (from Never to Often). All items on the BRIEF2 Parent form describe behaviors indicating executive function deficits. The internal consistency (n = 1,400; M age = 11.5 years; SD = 4.0 years), test-retest reliability (n = 163; M age = 11.3 years; SD = 4.1 years), and inter-rater reliability samples were derived from the norming sample (N = 1,400) for the Parent form, which was well-matched to United States population parameters for age, gender, parent education level, race/ethnicity, and geographic region. Internal consistency reliability estimates are coefficient alpha values, test-retest reliability estimates are Pearson correlation values stemming from administrations across an average of 2.9 weeks, and inter-rater reliability estimates are Pearson correlation values.
For all six BRIEF2 (Gioia et al., 2015a) parent subscales employed in this study, the mean internal consistency coefficients (across age groups) were above .84. Four of the subscales (e.g., Inhibit, Self-Monitor, Working Memory, and Plan/Organize) had mean test-retest reliability coefficients (across age groups) above .80, and two of the subscales (e.g., Shift and Emotional Control) had mean test-retest reliability coefficients below .80. Parent-to-parent inter-rater reliability was calculated for a nonclinical sample (n = 149; M age = 10.8; SD = 3.8) and a clinical sample (n = 287; M age = 11.4; SD = 3.4). Mean inter-rater reliability coefficients for most subscales employed in this study were above .70 (across age groups), with the exception of the Shift (r = .64) and the Self-Monitor (r = .57) subscales, for the nonclinical sample. The mean inter-rater reliability coefficients for most subscales employed in this study were above .60 (across age groups), except for the Plan/Organize (r = .43), Self-Monitor (r = .44), and Working Memory (r = .52) subscales, for the clinical sample. The total score, the Global Executive Composite, had a mean internal consistency reliability value of .97, a mean test-retest reliability coefficient of .88, and a mean inter-rater reliability coefficient of .86 for the nonclinical sample and .56 for the clinical sample.

Validity evidence supporting the BRIEF2 (Gioia et al., 2015a) parent subscales and total score include evidence based on content; items were grounded in theory and comprehensive domain sampling, which included extracting items from clinical work and adding items from experienced clinicians and establishing agreement across an expert panel of 12 pediatric neuropsychologists for each of the items included in the intended subscale (Gioia et al., 2015b). The BRIEF2 Professional Manual (Gioia et al., 2015b) also provided evidence based on internal structure (via item-total correlations between subscale items and total scores, intercorrelations between subscales, and exploratory and factor analyses), and evidence based on external
relations, including studies with 11 clinical groups, as well as studies comparing the BRIEF2 Parent form to other behavior rating scales and to tests of cognitive abilities.

The BRIEF2 (Gioia et al., 2015a) Self-Report form was developed for adolescents ages 11 to 18 years to rate their own problem behaviors that have occurred over the past 6 months. This form contains a total of 58 items that are rated on a 3-point scale (from Never to Often). All items on the BRIEF2 Self-Report form describe behaviors indicating executive function deficits. The internal consistency ($n = 803; M$ age $= 11.5$ years; $SD = 4.0$ years), test-retest reliability ($n = 190; M$ age $= 11.3$ years; $SD = 4.1$ years), and inter-rater reliability ($n = 472; M$ age $= 11.3$ years; $SD = 4.1$ years) samples were derived from the norming sample ($N = 803$) for the Self-Report form, which was well-matched to U.S. population parameters for age, gender, parent education level, race/ethnicity, and geographic region. Internal consistency reliability estimates are coefficient alpha values, test-retest reliability estimates are Pearson correlation values stemming from administrations across an average of 3.7 weeks, and inter-rater reliability estimates are Pearson correlation values.

For all six of the BRIEF2 (Gioia et al., 2015a) self-report subscales employed in this study, the mean internal consistency coefficients were above .84 (across age groups). All six of the subscales had mean test-retest reliability coefficients above .60, but below .80 (across age groups). Self-report-to-parent inter-rater reliability was calculated for a nonclinical sample ($n = 472; M$ age $= 14.6; SD = 2.3$) and a clinical sample ($n = 458; M$ age $= 14.7; SD = 2.0$). Mean inter-rater reliability coefficients for all subscales employed in this study were above .50 (across age groups), except for the Self-Monitor subscale ($r = .41$) for the nonclinical sample. The mean inter-rater reliability coefficients for most subscales employed in this study were above .30 (across age groups), with the exception of the Shift ($r = .22$) and Self-Monitor ($r = .18$).
subscales, for the clinical sample. The total score, the Global Executive Composite, had a mean internal consistency of .97, a mean test-retest reliability coefficient of .85, and a mean inter-rater reliability coefficient of .71 for the nonclinical sample and .25 for the clinical sample.

Validity evidence supporting the BRIEF2 (Gioia et al., 2015a) self-report subscales and total score include evidence based on item content and internal structure (via item-total correlations between subscale items and total scores, intercorrelations between subscales, and exploratory and factor analyses), and external relations via studies with 8 clinical groups as well as studies comparing the BRIEF2 Self-Report form to other behavior rating scales and to tests of cognitive abilities (Gioia et al., 2015b).

Comprehensive Executive Function Inventory (CEFI). The CEFI (Naglieri & Goldstein, 2013a) is a rating scale used to measure behaviors associated with executive functioning in children and adolescents ages 5 to 18 years. For the purpose of this study, only the CEFI Parent form and the CEFI Self-Report form were employed. In addition, only the scores derived from the seven subscales (Inhibitory Control, Flexibility, Self-Monitoring, Emotion Regulation, Planning, Organization, and Working Memory) and the Full Scale—all of which are common across the CEFI forms—were analyzed. All scores are standard scores ($M = 100; SD = 15$) based on combined gender norms for age group. For the CEFI Parent form, the combined gender norms are based on two age groups (i.e., norm blocks for ages 5 to 11 and 12 to 18). For the CEFI Self-Report form, combined gender norms based on one age group (i.e., norm block for ages 12 to 18). See Table 2 for a description of these scores.
### Table 2

**Description of CEFI Subscales and Full Scale**

<table>
<thead>
<tr>
<th>Index or Scale</th>
<th>Number of Items Parent</th>
<th>Number of Items Self-Report</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inhibitory Control</td>
<td>10</td>
<td>10</td>
<td>The ability to control impulsive behavior and to anticipate the consequences of behavior.</td>
</tr>
<tr>
<td>Flexibility</td>
<td>7</td>
<td>7</td>
<td>The ability to appropriately adapt behavior according to environmental demands and to generate alternative problem-solving approaches.</td>
</tr>
<tr>
<td>Self-Monitoring</td>
<td>10</td>
<td>10</td>
<td>The ability to evaluate and regulate behavior.</td>
</tr>
<tr>
<td>Emotion Regulation</td>
<td>9</td>
<td>9</td>
<td>The ability to regulate emotions and emotional responses to environmental stimuli.</td>
</tr>
<tr>
<td>Planning</td>
<td>11</td>
<td>11</td>
<td>The ability to strategize the necessary steps in order to accomplish goals, plan ahead, and to make good decisions.</td>
</tr>
<tr>
<td>Organization</td>
<td>10</td>
<td>10</td>
<td>The ability to effectively manage multiple tasks independently, organize thoughts, work systematically, and manage time.</td>
</tr>
<tr>
<td>Working Memory</td>
<td>11</td>
<td>11</td>
<td>The ability to hold information in mind in order to successfully complete multi-step instructions and to remember important dates and deadlines.</td>
</tr>
<tr>
<td>Full Scale</td>
<td>100</td>
<td>100</td>
<td>This score provides an estimate of a child’s or adolescent’s overall level of executive function across subscales.</td>
</tr>
</tbody>
</table>

The CEFI (Naglieri & Goldstein, 2013a) Parent form was developed for parents to rate behaviors that have occurred during the past 4 weeks for children ages 5 to 18 years. The CEFI Parent form contains a total of 100 items that are rated on a 6-point scale (from *Never* to *Always*). A total of 77 items describe behaviors indicating executive function deficits, and a total of 23 items describe behaviors indicating appropriate application of executive functions and are
worded positively. The internal consistency \((n = 676-698; \text{age group 12 to 18})\), test-retest reliability \((n = 171; M \text{age} = 11.7 \text{years}; SD = 4.0 \text{years})\) and inter-rater reliability \((n = 51; M \text{age} = 10.2 \text{years}; SD = 3.7 \text{years})\) samples were derived from the norming sample \((N = 1,396)\) for the Parent form. Notably, the sample reported in the study for the internal consistency of the Parent form is for age group 12 to 18 from the norming sample, whereas the other samples reported from the Parent form are for ages 5 to 18 from the norming sample. The norming sample was collected via a stratified sampling plan based on the 2009 U.S. Census. Age and gender were balanced across samples, and race/ethnicity, parent education level, and geographic region were matched to the 2009 U.S. Census data. Internal consistency reliability estimates are coefficient alpha values, test-retest reliability estimates are uncorrected Pearson correlation values stemming from administrations across an average of 17.1 days, and inter-rater reliability estimates are uncorrected Pearson correlation values.

For all seven CEFI (Naglieri & Goldstein, 2013a) parent subscales employed in this study, the mean internal consistency coefficients were above .85 (across age group 12 to 18). Most of the subscales had mean test-retest reliability coefficients above .84, except for the Flexibility subscale \((r = .79; \text{across age groups})\). Parent-parent mean inter-rater reliability coefficients were above .70 (across age groups) for five subscales (e.g., Inhibitory Control, Self-Monitoring, Organization, Working Memory, and Planning), but were below .70 (across age groups) for the Emotion Regulation and Flexibility subscales. The total score, the Full Scale, had a mean internal consistency of .97, a mean test-retest reliability coefficient of .89, and a mean inter-rater reliability coefficient of .83.

Validity evidence supporting the CEFI (Naglieri & Goldstein, 2013a) parent subscales and total score include evidence of internal structure via item-level and scale-level exploratory
factor analyses; congruence analyses to evaluate the similarity of the CEFI’s internal structure across several demographic groups (e.g., gender, age, race/ethnicity, and clinical/educational status); and analyses of covariance to examine the consistency of CEFI parent scores across different groups (e.g., race/ethnicity, population, and rater; Naglieri & Goldstein, 2013b).

Validity evidence based on external relations also stems from a study comparing the CEFI Parent form to the BRIEF (Gioia et al., 2000) Parent form as well as studies conducted with the CEFI Parent form and 10 clinical groups (Naglieri & Goldstein, 2013b).

The CEFI (Naglieri & Goldstein, 2013a) Self-Report form was developed for adolescents ages 12 to 18 years to rate their own problem behaviors that have occurred over the past 4 weeks. The CEFI Self-Report form contains a total of 100 items that are rated on a 6-point scale (from Never to Always). A total of 77 items describe behaviors indicating executive function deficits, and a total of 23 items describe behaviors indicating appropriate application of executive functions and are worded positively. The internal consistency, rater consistency (n = 126), and test-retest reliability (n = 200; M age = 15.2 years; SD = 1.7 years) samples were derived from the norming sample (N = 700) for the Self-Report form, which was collected via a stratified sampling plan based on the 2009 U.S. Census. Age and gender were balanced across samples, and race/ethnicity, parent education level, and geographic region were matched to the 2009 U.S. Census data. Internal consistency reliability estimates are coefficient alpha values, test-retest reliability estimates are uncorrected Pearson correlation values stemming from administrations across an average of 18 days, and rater consistency estimates are uncorrected Pearson correlation values. Of note, the CEFI Technical Manual (Naglieri & Goldstein, 2013b) provided consistency estimates between rater types for parent-to-self-report for all six subscales used in the analyses, but not for the total score.
Internal consistency for the CEFI (Naglieri & Goldstein, 2013a) self-report subscales was calculated for a nonclinical sample ($n = 667-700$) and a clinical sample ($n = 148-205$). The mean internal consistency coefficients for most subscales employed in this study were above .80 (across age groups), with the exception of the Emotion Regulation ($r = .78$), Self-Monitoring ($r = .78$), and Flexibility ($r = .77$) subscales, for the nonclinical sample. The mean internal consistency coefficients for most of the subscales employed in this study were above .80 (across age groups), except for the Flexibility ($r = .72$) and Self-Monitoring ($r = .74$) subscales, for the clinical sample. Three subscales (e.g., Flexibility, Organization, and Planning) had mean test-retest reliability coefficients above .80 (across age groups), and four subscales (e.g., Emotion Regulation, Self-Monitoring, Inhibitory Control, and Working Memory) had mean test-retest reliability coefficients below .80. For all seven subscales employed in this study, the mean rater consistency coefficients for parent-adolescent raters were above .38 but below .70 (across age groups; Naglieri & Goldstein, 2013b). The total score, the Full Scale, had a mean internal consistency coefficient of .97 and a mean test-retest reliability coefficient of .78 for the nonclinical and clinical samples.

Validity evidence supporting the CEFI (Naglieri & Goldstein, 2013a) self-report subscales and total score include evidence of internal structure via item-level and scale-level exploratory factor analyses; congruence analyses to evaluate the similarity of the CEFI’s internal structure across several demographic groups (e.g., gender, age, race/ethnicity, and clinical/educational status); and analyses of covariance to examine the consistency of CEFI self-report scores across different groups (e.g., race/ethnicity, population, and rater; Naglieri & Goldstein, 2013b). Validity evidence based on external relations also stems from a study comparing the CEFI Self-Report form to the BRIEF (Gioia et al., 2000) Self-Report form as well
as studies conducted with the CEFI Self-Report form and 5 clinical groups (Naglieri & Goldstein, 2013b).

**Procedures**

**Recruitment and selection of participants.** Parents of adolescents enrolled in sixth through twelfth grades in a rural public school district located in the south-eastern region of the United States were recruited in winter 2018 of the academic year via a Letter of Invitation (see Appendix B), which included an explanation of the study. Sealed envelopes containing the Letter of Invitation, an Informed Consent Form (see Appendix C) for parents to keep for their records, an Informed Consent Form for parents to sign and return, and a Demographics Information Form (see Appendix D) were distributed by designated school teachers and administrators from selected schools to adolescents to send home to their caregivers. Inclusion criteria for the study included adolescents ages 12 to 17 currently enrolled in general education and special education at the select middle and high school. Parents were asked to return the signed Informed Consent Form and the Demographics Information Form to the researcher or designated school administrator in a sealed envelope. Selected school administrators and teachers provided the unopened envelopes from parents to the researchers. Overall, forms were sent home to parents of 250 adolescents, and 56 parents (22%) returned forms permitting their children to participate.

**Post-consent contact and scheduling of rating scale administration.** After parental consent was obtained, the researcher provided a list of participants whose parents provided consent to participate in the study to designated school administrators and teachers in order to facilitate contact with these participants. The researcher coordinated with school administrators to schedule times to administer the rating scales to adolescents during school hours (8:00 am to 2:30 pm). All participant information was maintained in a secure filing cabinet by the researcher.
Under no circumstances were specific results of the study provided to parents, school administrators, teachers, or adolescents.

**Parent administration.** After parents of adolescents from selected schools consented to participate in the study by returning their materials to designated school administrators and teachers, the parent participants were assigned to one of the two counterbalanced conditions (addressing scale order) based on blocked random assignment (Keppel & Wickens, 2004). The rating scales were sent home with the participating adolescents in a packet containing the BRIEF2 (Gioia et al., 2015a) and the CEFI (Naglieri & Goldstein, 2013a) parent rating scales and instructions regarding how the parent should complete the rating scales (see Appendix F). Parents were instructed to complete the rating scales within 1 week of receiving them and to place the completed rating scales in an envelope provided by the researcher, seal the envelope, and return the sealed envelope to the designated school administrators or teachers. The school administrators and teachers were instructed to place the packet in the researcher’s school mailbox or to give the packet directly to the researcher. All completed materials were maintained in a secure filing cabinet.

If parents did not return the rating scales within 1 week of delivery, they were contacted via phone by the researcher to remind them to complete and return the forms; parents were reminded by the researcher via phone a total of four attempts. If parents were unsuccessfully contacted via phone, an additional packet containing the rating scales was sent home with adolescents to give to their parents to complete and return within 1 week of delivery. Of parents who consented to participate in the study, a total of 42 out of 56 parents (75%) completed and returned the packet. Forty-eight percent of parents completed the forms within 1 week of delivery, 17% of parents completed the forms within 2 weeks of delivery, 26% of parents
completed the forms within 1 month of delivery, and 7% of parents completed the forms within 3 months of delivery.

A brief hand-out with evidence-based study and organizational strategies for adolescents was distributed to parents who completed the study after the study closed in May 2018; the handout was sent home with the adolescent in an envelope with the parent’s name on it. Additionally, participants who completed the study entered their name into a raffle to win a $15.00 Amazon gift card. There were a total of 10 $15.00 gift cards. Gift cards were distributed via mail to winning participants following the completion of the study.

The BRIEF2 Professional Manual (Gioia et al., 2015b) indicates that the BRIEF2 Parent form cannot be scored if there are more than 12 missing responses on the form. If there were fewer than 12 missing responses on the BRIEF2 Parent form, the form was scored and the data were utilized in the study, as this situation may reflect what happens in real life. If only 1 item contributing to a BRIEF2 parent subscale was missing, it was assigned a score of 1 (Never). If 5 or fewer items on the BRIEF2 Parent form were missing responses, the parent was contacted via phone by the researcher in which parents were read the items and asked to respond. If more than 5 items on the BRIEF2 Parent form were missing responses, the form (with missing items highlighted) was mailed to the parent to complete and return. If more than 12 items were missing responses and the parent refused to complete the missing items, was unsuccessfully contacted after 3 attempts, or did not complete and return the form within 1 week of delivery, the data were not utilized. If responses to items 18, 36, and 54, were missing on the BRIEF2 Parent form, they were not remedied because they do not affect the calculation of the subscale and total scores (Gioia et al., 2015b). A total of 6 cases (14%) had 1 item missing from the BRIEF2 Parent forms.
and were unable to be remedied via the procedures outlined above. Therefore, the missing item for each case was scored as a 1 (Gioia et al., 2015b).

The CEFI Technical Manual (Naglieri & Goldstein, 2013b) indicates that the CEFI Parent form yields a prorated raw score for the total score if fewer than 6 responses are missing across items contributing to the total score, and a prorated raw score for the subscales, if only 1 response is missing across items contributing to a subscale score. If fewer than 6 items were missing responses on the CEFI Parent form, the parent was contacted via phone by the researcher in which they were read the items and asked to respond. If 6 or more items on the CEFI Parent form were missing responses, the form (with missing responses highlighted) was mailed to the parent to complete and return. Prorated raw scores for the CEFI Parent form were only utilized if the parent refused to complete missing items, was unsuccessfully contacted after 3 attempts, or did not complete and return the form within 1 week of delivery. A total of 9 (21%) cases contain prorated scores.

Adolescent administration. Prior to administering the rating scales in the school setting, assent was obtained from adolescents whose parents provided consent to participate in the study. Assent was considered obtained if adolescents read and signed the Assent Form (see Appendix E). For adolescents who provided assent, the rating scales were presented in a counterbalanced order; they were assigned to one of two conditions reflecting the order of administration (Keppel & Wickens, 2004). The rating scales were administered by the researcher to adolescents individually or in a group format, and administration lasted 30 minutes or less. Adolescents were instructed to place the completed rating scales in an envelope provided by the researcher, seal the envelope, and return the sealed envelope to the researcher. Of the consented adolescents, 100% provided assent. All of the rating scales were administered during school hours and during non-
academic activities (e.g., recess, computer lab, drafting, and gym), and 69% of rating scales were administered in a group format. When scores from independent and group administrations were contrasted, there were no statistically significant differences across the total scores for the BRIEF2 Self-Report form, \( t(41) = 2.36, p = .13 \) and the CEFI Self-Report form, \( t(41) = 2.14, p = .15 \). Thus, no confounds related to setting on the participants’ total scores were evident.

The *BRIEF2 Professional Manual* (Gioia et al., 2015b) indicates that the BRIEF2 Self-Report form cannot be scored if there are more than 10 missing responses to items on the form. In order to prevent the likelihood of missing responses on the BRIEF2 Self-Report form, the form was screened for missing responses after each adolescent returned the form to the researcher, and the adolescent was asked to complete items with missing responses before being dismissed. If fewer than 10 responses on the BRIEF2 Self-Report form were missing, the form was scored and the data were utilized, as this situation may reflect what happens in real life. If 5 or fewer items on the BRIEF2 Self-Report form were missing responses, the missing items were highlighted and the adolescent was asked to complete the items. If only 1 item contributing to a subscale was missing a response, the adolescent refused to complete missing responses, or the adolescent was unsuccessfully contacted to complete missing responses after four attempts, the item was assigned a score of 1 (*Never*; Gioia et al., 2015b). If responses to items 18, 36, and 54 were missing on the BRIEF2 Self-Report form, they were not remedied because they do not affect the calculation of the subscale and total scores (Gioia et al., 2015b). A total of 4 cases (9%) had 1 item missing from the BRIEF2 Self-Report forms and were unable to be remedied via the procedures outlined above. Therefore, the missing item for each case was scored as a 1.

The *CEFI Technical Manual* (Naglieri & Goldstein, 2013b) indicates that the CEFI Self-Report form yields a prorated raw score for the total score if there are fewer than 6 missing
responses across items contributing to the total score and a prorated raw score for the subscales if there is only 1 missing response contributing to a subscale score. After each adolescent returned the CEFI Self-Report rating scale to the researcher, the form was screened for missing responses before the adolescent was dismissed. If 1 or more items on the CEFI Self-Report form were missing responses, the items were highlighted and the adolescent was asked to complete the items. Prorated raw scores for the CEFI Self-Report form were only utilized if the adolescent refused to complete missing items or was unsuccessfully contacted after four attempts. A total of 9 (21%) cases contain prorated scores.

**BRIEF2 Parent and Self-Report forms and the CEFI Parent and Self-Report forms were independently scored by one of three, second-year graduate students in school psychology. The graduate students passed two graduate-level assessment courses and a graduate-level assessment practicum, and two of the three graduate students were specifically trained in psychometrics. Additionally, all three graduate students completed one training session with the researcher prior to scoring any of the rating scales. Training sessions consisted of reviewing administration and scoring procedures for the rating scales included in the study as well as direct instruction in how to score BRIEF2 and CEFI Parent and Self-Report rating scales. The graduate students and the researcher obtained norm-referenced standardized scores for the BRIEF2 and the CEFI rating scales and norm-referenced T Scores for the CEFI rating scales using norms tables included in rating scale manuals. During the training session, each graduate student submitted scored BRIEF2 and CEFI Parent and Self-Report rating scales in order to ensure competence (i.e., fewer than four scoring errors across all four rating forms) in scoring. The protocols were reviewed by the researcher to ensure that no scoring errors were present, and minor errors were discussed with the scorers during the training session. Each graduate student demonstrated competency in**
scoring all rating scales prior to scoring the rating scales for the study. The graduate students completed the scoring within two weeks of rating scale administration. After scoring each rating scale, the graduate students placed the rating scales in folders in a filing cabinet that was maintained by the researcher. The researcher reviewed 100% of the rating scales for scoring errors before entering the scores into the database.

**Results**

**Data Screening and Tests of Assumptions**

All study analyses were conducted with the IBM Statistical Package for the Social Sciences (IBM SPSS) Statistics 24.0 (IBM Corp, 2016). Preliminary data analyses were conducted using the distribution of scores for each of the two instruments for each rater to ensure that the assumptions of multivariate analysis and correlations were not violated (Tabachnick & Fidell, 2013). Given that the BRIEF2 (Gioia et al., 2015a) yields T Scores and the CEFI (Naglieri & Goldstein, 2013a) yields standard scores, the standard scores from the CEFI total score and subscale scores were transformed into z scores and subsequently into T Scores; the resulting T Scores were then reversed scaled in order to be consistent with the BRIEF2 scaling. Results revealed that there were no missing values across instruments. There was a total of 2 univariate outliers across the BRIEF2 Parent and Self-Report rating scales, and 1 univariate outlier across the CEFI Parent and Self-Report rating scales (z score < |3.00|). Specifically, there was 1 univariate outlier for the BRIEF2 Parent rating scale (Inhibition z score = 3.05) and 1 univariate outlier for the BRIEF2 Self-Report rating scale (Plan/Organize z score = 3.24). Additionally, there was 1 univariate outlier for the CEFI Self-Report rating scale (Planning z score = 3.07); the CEFI Parent rating scale did not evidence any univariate outliers. The two cases across the BRIEF2 Parent and Self-Report rating scales and the one case from the CEFI
Self-Report rating scale were maintained in order to maintain the authenticity of the
generalizability study. No multivariate outliers were found across the BRIEF2 and CEFI parent
and self-report variables.

Descriptive statistics for all total scores and subscale scores revealed score means
generally consistent with the population mean, according to normative data from each
instrument. Similarly, total score and subscale score standard deviation values were generally
consistent with the population mean, as restriction and expansion of range in norm-referenced
total scores and subscale scores were not evident (e.g., standard deviation values smaller or
larger than population values, respectively). All univariate skewness and kurtosis values for total
scores were less than │1.0│. The majority of univariate skewness and kurtosis values for
subscale scores were less than │1.0│ except for five subscales across the BRIEF2 (Gioia et al.,
2015a) Parent and Self-Report rating scales and one subscale from the CEFI (Naglieri &
Goldstein, 2013a) Self-Report rating scales; the univariate skewness and kurtosis values for
subscale scores from the CEFI Parent rating scales were within normal limits. Non-normality
was noted for the BRIEF2 Self-Report Plan/Organize subscale (skewness = 1.27 and kurtosis =
2.14); the BRIEF2 Parent Inhibition (skewness = 1.03), Emotional Control (skewness = 1.03),
Self-Monitor (kurtosis = -1.26), and Plan/Organize (kurtosis = -1.05) subscales; and the CEFI
Self-Report Planning subscale (kurtosis = 1.36). No skewness or kurtosis values were extreme
(Tabachnick & Fidell, 2013). All other assumptions of paired-samples t-tests were judged not to
be violated.

Table 3 includes the means and standard deviations for each instrument, total score, and
subscale score by rater. The BRIEF2 (Gioia et al., 2015a) means ranged from 49.12
(Plan/Organize) to 53.02 (Shift) for the parent raters and from 54.36 (Plan/Organize) to 57.71
(Working Memory) for the adolescent raters. The standard deviations for the parent raters ranged from 9.13 (Plan/Organize) to 10.47 (Shift). The standard deviations were less than 10 for the GEC, Inhibit, Self-Monitor, Emotional Control, and Plan/Organize scales and greater than 10 for the Shift and Working Memory scales, which indicates restriction and expansion of range for the parent sample, respectively. The standard deviations ranged from 8.84 (Shift) to 11.30 (Emotional Control) for the adolescent raters. The standard deviations were less than 10 for the GEC, Shift, and Plan/Organize scales and greater than 10 for the Inhibit, Self-Monitor, Emotional Control, and Working Memory scales, which indicates restriction and expansion of range for the adolescent sample, respectively. The CEFI (Naglieri & Goldstein, 2013a) means ranged from 47.02 (Self-Monitoring) to 48.67 (Inhibitory Control) for the parent raters and from 48.71 (Flexibility) to 52.68 (Planning) for the adolescent raters. The standard deviations ranged from 8.27 (Organization) to 10.23 (Emotion Regulation) for the parent raters. The standard deviations were less than 10 for four of the seven scales for the parent raters, including the Full Scale, Self-Monitoring, Planning, and Organization scales; the standard deviations were greater than 10 for the Inhibitory Control, Flexibility, Emotion Regulation, and Working Memory scales. The standard deviations for the adolescent raters ranged from 7.75 (Full Scale) to 9.62 (Working Memory); all standard deviations were less than 10 for the adolescent raters.

**Inter-rater Reliability and Mean Differences across Raters**

To examine the inter-rater reliability of the total scores and subscale scores, one correlation was calculated between the total scores and subscale scores from both raters from each instrument, resulting in a total of 15 correlations. On the right side of Table 3, uncorrected correlations across raters, mean differences across raters, and the results of independent-samples $t$-tests for each instruments’ total scores and subscale scores are presented. The following general
labels for correlations were used for this study: negligible, .00 to .19; weak, .20 to .39; moderate, .40 to .69; strong, .70 to .89; and very strong, .90 to 1.0 (Floyd et al., 2008). Inter-rater reliability coefficients for the BRIEF2 (Gioia et al., 2015a) ranged from .07 (Shift) to .35 (Emotional Control) and inter-rater reliability coefficients for the CEFI (Naglieri & Goldstein, 2013a) ranged from .04 (Self-Monitor and Organization) to .42 (Inhibitory Control). Inter-rater reliability correlations ranged from negligible to weak, for the BRIEF2 total score and subscale scores. Specifically, the BRIEF2 GEC evidenced a weak inter-rater reliability correlation. Similarly, four of the BRIEF2 subscales (Inhibit, Emotional Control, Working Memory, and Plan/Organize) also yielded weak inter-rater reliability correlations. Notably, two BRIEF2 subscales (Shift and Self-Monitor) yielded negligible inter-rater reliability correlations. Commensurate with the BRIEF2, the CEFI (Naglieri & Goldstein, 2013a) evidenced positive inter-rater reliability correlations, ranging from negligible to moderate, across the total score and subscale scores (see Table 3). The CEFI Full Scale yielded weak inter-rater reliability correlations; weak inter-rater reliability correlations also were found for four of the CEFI subscales (Flexibility, Emotion Regulation, Working Memory, and Planning). Two of the CEFI subscales evidenced negligible inter-rater reliability correlations (Self-Monitoring and Organization). Of note, a moderate inter-rater reliability correlation was found for the Inhibitory Control subscale. Mean differences across total scores, subscale scores, and raters were approximately 6 standard score points or less, and all t-tests revealed nonsignificant mean differences between total scores and subscale scores, ps > .001, across raters. Hedges’ $g$ suggested a weak effect for the Inhibit (.47), Shift (.33), Self-Monitor (.43), and Emotional Control (.35) scales and a moderate effect for the GEC (.64), Working Memory (.57), and Plan/Organize (.58) scales across raters for the BRIEF2. Hedges’ $g$ suggested a weak effect for
the Full Scale (.44), Inhibitory Control (.35), Flexibility (.05), Emotion Regulation (.30), Working Memory (.39), and Organization (.35) scales and a moderate effect for the Self-Monitoring (.50) and Planning (.60) scales across raters for the CEFI.
Table 3

Descriptive Statistics by Instrument across Rater, and Inter-Rater Reliability Correlations for Total Scores and Subscale Scores

<table>
<thead>
<tr>
<th>Total score/ subscale score</th>
<th>Rater</th>
<th>Rater</th>
<th>Inter-Rater Reliability Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parent</td>
<td>Adolescent</td>
<td>r</td>
</tr>
<tr>
<td><strong>BRIEF2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEC</td>
<td>50.79</td>
<td>9.24</td>
<td>39-73</td>
</tr>
<tr>
<td>Inhibit</td>
<td>50.55</td>
<td>9.31</td>
<td>38-79</td>
</tr>
<tr>
<td>Shift</td>
<td>53.02</td>
<td>10.47</td>
<td>40-83</td>
</tr>
<tr>
<td>Self-Monitor</td>
<td>51.26</td>
<td>9.41</td>
<td>39-70</td>
</tr>
<tr>
<td>Emotional Control</td>
<td>50.83</td>
<td>9.86</td>
<td>40-74</td>
</tr>
<tr>
<td>Working Memory</td>
<td>51.79</td>
<td>10.14</td>
<td>39-79</td>
</tr>
<tr>
<td>Plan/Organize</td>
<td>49.12</td>
<td>9.13</td>
<td>37-66</td>
</tr>
<tr>
<td><strong>CEFI</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full Scale</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inhibitory Control</td>
<td>48.67</td>
<td>10.04</td>
<td>29-69</td>
</tr>
<tr>
<td>Flexibility</td>
<td>48.19</td>
<td>10.06</td>
<td>23-65</td>
</tr>
<tr>
<td>Self-Monitoring</td>
<td>47.02</td>
<td>8.93</td>
<td>29-67</td>
</tr>
<tr>
<td>Emotion Regulation</td>
<td>48.24</td>
<td>10.23</td>
<td>31-70</td>
</tr>
<tr>
<td>Working Memory</td>
<td>48.05</td>
<td>10.04</td>
<td>30-70</td>
</tr>
<tr>
<td>Planning</td>
<td>47.10</td>
<td>9.66</td>
<td>28-66</td>
</tr>
<tr>
<td>Organization</td>
<td>47.97</td>
<td>8.27</td>
<td>29-65</td>
</tr>
</tbody>
</table>
Table 3 (Continued)

Note. Composite and subscale scores are age-based T scores ($M = 50$, $SD = 10$) unless otherwise noted. GEC = Global Executive Composite. BRIEF2 = Behavior Rating Inventory of Executive Function, Second Edition; CEFI = Comprehensive Executive Functioning Inventory.

*p < .05 (two-tailed).

**p < .001 (two-tailed).
Convergent Validity and Mean Differences across Instruments

To examine the convergent validity evidence supporting the total scores and subscale scores, correlations within and across instruments for each set of raters were conducted (see Tables 4 and 5). First, intercorrelations within each instrument for both raters were conducted, resulting in seven correlations. The BRIEF2 (Gioia et al., 2015a) yielded strong positive intercorrelations between the GEC and all six subscale scores. Intercorrelations for the BRIEF2 subscale scores were positive and ranged from moderate to strong, although the majority of intercorrelations were moderate. The CEFI (Naglieri & Goldstein, 2013a) evidenced moderate to very strong positive intercorrelations between the Full Scale and the seven subscale scores. Intercorrelations for the CEFI subscale scores were positive and ranged from weak to strong. Similar to the BRIEF2, the majority of the intercorrelations across the CEFI subscales were moderate. Of note, for both instruments, the strength of the intercorrelations generally varied across raters.

With regard to rater, adolescent raters evidenced a strong positive correlation between the BRIEF2 (Gioia et al., 2015a) Self-Report GEC and all of the BRIEF2 Self-Report subscales. Adolescent raters evidenced positive correlations within the moderate range across subscale scores on the BRIEF2 Self-Report. Although the majority of subscale score correlations were moderate for adolescent raters, there was one strong correlation (between Working Memory and Plan/Organize). Similarly, there were moderate (between Full Scale and Flexibility) to very strong (between Full Scale and Self-Monitor and Full Scale and Planning) positive correlations between the CEFI (Naglieri & Goldstein, 2013a) Full Scale and the CEFI subscale scores for the adolescent raters. Adolescent raters yielded 17 moderate correlations and three strong
correlations between the CEFI Self-Report subscales; however, there was one weak correlation between the Inhibitory Control and Organization subscales.

Consistent with the adolescent raters, the parent raters yielded a strong positive correlation between the BRIEF2 (Gioia et al., 2015a) Parent GEC and all of the BRIEF2 Parent subscales. Furthermore, parent raters yielded 10 moderate correlations and five strong correlations for the BRIEF2 Parent subscale scores. Parent raters evidenced strong to very strong (between Full Scale and Self-Monitor and Full Scale and Planning) positive correlations between the CEFI (Naglieri & Goldstein, 2013a) Parent Full Scale and the CEFI Parent subscale scores. There were eight moderate correlations, 12 strong correlations, and one very strong correlation (between Flexibility and Self-Monitoring) across CEFI Parent subscales.

Intercorrelations across instrument, as produced by each rater, were also conducted in order to examine convergent validity across total scales and subscales developed to measure the same construct (see Tables 4 and 5). The scales of interest included the following six pairs of subscales hypothesized to be theoretically related: the BRIEF2 (Gioia et al., 2015a) GEC and the CEFI (Naglieri & Goldstein, 2013a) Full Scale; the BRIEF2 Inhibit and the CEFI Inhibitory Control subscales; the BRIEF2 Shift and the CEFI Flexibility subscales; the BRIEF2 Emotional Control and the CEFI Emotion Regulation subscales; the BRIEF2 Plan/Organize and the CEFI Planning and CEFI Organization subscales; and the Self-Monitor and Working Memory subscales from both instruments.

Intercorrelations revealed weak to moderate positive correlations across the BRIEF2 (Gioia et al., 2015a) and CEFI (Naglieri & Goldstein, 2013a) self-report rating scales. Specifically, a moderate correlation was found between the total scores from the BRIEF2 Self-Report (GEC) and the CEFI Self-Report (Full Scale) for the adolescent raters. Furthermore,
weak correlations were found between the BRIEF2 and CEFI self-report subscales, including Inhibit and Inhibitory Control, Shift and Flexibility, and Self-Monitor and Self-Monitoring. Moderate correlations were evidenced between the BRIEF2 and CEFI self-report Emotional Control and Emotion Regulation subscales and the Plan/Organize and Organizing subscales, as well as the Working Memory subscales from both self-report instruments.

Similar results were found across the BRIEF2 (Gioia et al., 2015a) Parent and CEFI (Naglieri & Goldstein, 2013a) Parent total scores and subscale scores for the parent raters. Moderate correlations were found for the total scores for the BRIEF2 Parent (GEC) and the CEFI Parent (Full Scale) rating scales. Additionally, moderate correlations were found between all subscale scores of interest across the BRIEF2 and CEFI parent rating scales, including Inhibit and Inhibitory Control, Shift and Flexibility, Emotional Control and Emotion Regulation, Plan/Organize and Planning and Organization, Self-Monitor and Self-Monitoring, and the Working Memory subscales from both instruments.

Several paired samples t-tests were conducted to compare the effects of each instrument on mean total score and subscale scores for each rater. Results revealed a significant effect of instrument on total scores, $t(41) = 4.14, p < .001$ for the adolescent raters. Specifically, adolescent raters scored higher on the BRIEF2 (Gioia et al., 2015a) Self-Report total score ($M = 56.79, SD = 9.44$) than on the CEFI (Naglieri & Goldstein, 2013a) Self-Report total score ($M = 51.71, SD = 7.75$), and Cohen’s $d$ (0.59) suggested a moderate effect. With regard to subscale scores, a significant effect of instrument on subscale scores was found for the Shift and Flexibility subscale scores, $t(41) = 4.45, p < .001$, and for the Working Memory subscale scores, $t(41) = 4.68, p < .001$, for the adolescent raters. Adolescent raters’ scores for the BRIEF2 Shift subscale ($M = 56.21, SD = 8.84$) were higher in comparison to the CEFI Flexibility subscale ($M$
= 48.71, SD = 9.19). In addition, adolescent raters scored the BRIEF2 Working Memory subscale (M = 57.71, SD = 10.45) higher than the CEFI Working Memory subscale (M = 51.94, SD = 9.62). Cohen’s d (0.83) suggested a strong effect for the Shift and Flexibility subscales, and Cohen’s d (0.57) for the Working Memory subscales suggested a moderate effect for the adolescent raters. For the parent raters, there was no significant effect of instrument on total scores, t(41) = 2.30, p = .03.
Table 4

*Intercorrelations across Instrument within Rater for Adolescents*

<table>
<thead>
<tr>
<th>BRIEF2</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. GEC</td>
<td>--</td>
<td>.71</td>
<td>.82</td>
<td>.73</td>
<td>.82</td>
<td>.84</td>
<td>.84</td>
<td>.59</td>
<td>.42</td>
<td>.35</td>
<td>.33</td>
<td>.50</td>
<td>.56</td>
<td>.51</td>
<td>.56</td>
</tr>
<tr>
<td>2. Inhibit</td>
<td>--</td>
<td>.52</td>
<td>.50</td>
<td>.63</td>
<td>.41</td>
<td>.47</td>
<td>.39</td>
<td>.34</td>
<td>.12</td>
<td>.17</td>
<td>.51</td>
<td>.32</td>
<td>.37</td>
<td>.37</td>
<td>.33</td>
</tr>
<tr>
<td>3. Shift</td>
<td>--</td>
<td>.54</td>
<td>.63</td>
<td>.60</td>
<td>.68</td>
<td>.38</td>
<td>.20</td>
<td>.27</td>
<td>.21</td>
<td>.35</td>
<td>.39</td>
<td>.24</td>
<td>.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Self-Monitor</td>
<td>--</td>
<td>.58</td>
<td>.61</td>
<td>.50</td>
<td>.36</td>
<td>.39</td>
<td>.31</td>
<td>.29</td>
<td>.30</td>
<td>.23</td>
<td>.41</td>
<td>.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Emotional Control</td>
<td>--</td>
<td>.56</td>
<td>.59</td>
<td>.48</td>
<td>.50</td>
<td>.35</td>
<td>.25</td>
<td>.61</td>
<td>.36</td>
<td>.34</td>
<td>.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Working Memory</td>
<td>--</td>
<td>.70</td>
<td>.54</td>
<td>.27</td>
<td>.11</td>
<td>.20</td>
<td>.37</td>
<td>.69</td>
<td>.45</td>
<td>.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Plan/Organize</td>
<td>--</td>
<td>.50</td>
<td>.28</td>
<td>.29</td>
<td>.31</td>
<td>.24</td>
<td>.48</td>
<td>.46</td>
<td>.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CEFI</th>
<th>8. Full Scale</th>
<th>--</th>
<th>.75</th>
<th>.67</th>
<th>.75</th>
<th>.75</th>
<th>.90</th>
<th>.78</th>
<th>.80</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. Inhibitory Control</td>
<td>--</td>
<td>.44</td>
<td>.45</td>
<td>.79</td>
<td>.57</td>
<td>.66</td>
<td>.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Flexibility</td>
<td>--</td>
<td>.54</td>
<td>.40</td>
<td>.44</td>
<td>.59</td>
<td>.42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Self-Monitor</td>
<td>--</td>
<td>.48</td>
<td>.41</td>
<td>.72</td>
<td>.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Emotion Regulation</td>
<td>--</td>
<td>.53</td>
<td>.59</td>
<td>.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Working Memory</td>
<td>--</td>
<td>.65</td>
<td>.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Planning</td>
<td>--</td>
<td>.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Organization</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Pearson product-moment correlation coefficients for the BRIEF2 are presented below the diagonal, and correlations for the CEFI are reported above the diagonal. We also recognize that there is no set standard for providing nominal labels for *r* values. GEC =
Table 4 (Continued)

Global Executive Composite. BRIEF2 = Behavior Rating Inventory of Executive Function, Second Edition; CEFI = Comprehensive Executive Functioning Inventory. All correlations significant at $p < .001$ (two-tailed) unless noted.

* $p < .05$ (two-tailed).

<sup>a</sup> correlation was nonsignificant.
Table 5

*Intercorrelations across Instrument within Rater for Parents*

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BRIEF2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. GEC</td>
<td>--</td>
<td>.80</td>
<td>.85</td>
<td>.73</td>
<td>.78</td>
<td>.89</td>
<td>.86</td>
<td>.61</td>
<td>.59</td>
<td>.45</td>
<td>.50</td>
<td>.54</td>
<td>.54</td>
<td>.54</td>
<td>.53</td>
</tr>
<tr>
<td>2. Inhibit</td>
<td>--</td>
<td>.65</td>
<td>.64</td>
<td>.83</td>
<td>.71</td>
<td>.57</td>
<td>.45</td>
<td>.55</td>
<td>.33*</td>
<td>.31*</td>
<td>.56</td>
<td>.35*</td>
<td>.36*</td>
<td>.32*</td>
<td></td>
</tr>
<tr>
<td>3. Shift</td>
<td>--</td>
<td>.55</td>
<td>.67</td>
<td>.66</td>
<td>.75</td>
<td>.54</td>
<td>.53</td>
<td>.49</td>
<td>.55</td>
<td>.45</td>
<td>.53</td>
<td>.40</td>
<td>.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Self-Monitor</td>
<td>--</td>
<td>.52</td>
<td>.63</td>
<td>.73</td>
<td>.51</td>
<td>.53</td>
<td>.39*</td>
<td>.43</td>
<td>.40</td>
<td>.38*</td>
<td>.51</td>
<td>.38*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Emotional Control</td>
<td>--</td>
<td></td>
<td>.71</td>
<td>.50</td>
<td>.38*</td>
<td>.52</td>
<td>.27*</td>
<td>.26*</td>
<td>.60</td>
<td>.26*</td>
<td>.28*</td>
<td>.27*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Working Memory</td>
<td>--</td>
<td></td>
<td>.66</td>
<td>.65</td>
<td>.59</td>
<td>.46</td>
<td>.56</td>
<td>.41</td>
<td>.68</td>
<td>.56</td>
<td>.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Plan/Organize</td>
<td>--</td>
<td></td>
<td>.49</td>
<td>.44</td>
<td>.33*</td>
<td>.36*</td>
<td>.52</td>
<td>.38*</td>
<td>.42</td>
<td>.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CEFI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Full Scale</td>
<td>--</td>
<td></td>
<td>.86</td>
<td>.89</td>
<td>.93</td>
<td>.76</td>
<td>.88</td>
<td>.95</td>
<td>.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Inhibitory Control</td>
<td>--</td>
<td></td>
<td>.72</td>
<td>.75</td>
<td>.82</td>
<td>.69</td>
<td>.77</td>
<td>.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Flexibility</td>
<td>--</td>
<td></td>
<td>.90</td>
<td>.63</td>
<td>.88</td>
<td>.75</td>
<td>.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Self-Monitor</td>
<td>--</td>
<td></td>
<td>.62</td>
<td>.83</td>
<td>.88</td>
<td>.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Emotion Regulation</td>
<td>--</td>
<td></td>
<td>.56</td>
<td>.66</td>
<td>.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Working Memory</td>
<td>--</td>
<td></td>
<td>.82</td>
<td>.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Planning</td>
<td>--</td>
<td></td>
<td>.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Organization</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5 (Continued)

*Note.* Pearson product-moment correlation coefficients for the BRIEF2 are presented below the diagonal, and correlations for the CEFI are reported above the diagonal. We also recognize that there is no set standard for providing nominal labels for $r$ values. GEC = Global Executive Composite. BRIEF2 = Behavior Rating Inventory of Executive Function, Second Edition; CEFI = Comprehensive Executive Functioning Inventory.

All correlations significant at $p < .001$ (two-tailed) unless noted.

* $p < .05$ (two-tailed).

$^a$ correlation was nonsignificant.
Generalizability Theory Analyses

Several univariate GT analyses were employed to examine dependability of the BRIEF2 (Gioia et al., 2015a) and the CEFI (Naglieri & Goldstein, 2013a) total scores and subscale scores across rater, instrument, and dimension (see Tables 6, 7, and 8). The first GT analysis (see Table 6) was conducted to assess the variance in total scores attributed to rater (parent and adolescent) and instrument (BRIEF2 and CEFI). This GT analysis employed a two-facet model design, with rater and instrument specified as the two facets and as random effects because the aim of this analysis was to generalize similar constructs from a universe of possible indicators of each construct. The total scores were entered into a GT analysis to examine their dependability.

Variance components were calculated to provide overall indexes of dependability (Brennan, 2001; Shavelson & Webb, 1991). The variance estimate attributable to differences in total scores was considered the universal score variance; it was used as the numerator in the formula to calculate the dependability coefficients. The variance estimates attributable to rater, instrument, all interactions, and to residual (i.e., unexplained) variance, was divided by the number of variations associated with each facet, resulting in error variance. The denominator of the formula included the sum of the universal score variance and error variance.

Table 6 provides the variance component estimates for the subject, the rater, the instrument, and for all associated interactions, on total scores. For reference, the object of measurement, variance attributable to individual differences across subjects, accounted for 15% of variance; therefore, the remainder of the variance was due to systematic or random error. The largest proportion of error variance was attributed to the subject-by-rater interaction; it accounted for 33% of variance. The second largest proportion of variance was attributed to the rater facet, which contributed 11% of variance. The instrument facet (7%), subject-by-instrument
interaction (3%), and rater-by-instrument interaction (1%) contributed minimal variance in total scores. Residual variance was 29%. The dependability coefficient for all facets was .30, which indicates weak dependability.

Table 6

Variance Component Estimates and Absolute Dependability Coefficients by Score Comparison

<table>
<thead>
<tr>
<th>Facet</th>
<th>Total Score</th>
<th>Percent of variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>14.804</td>
<td>15%</td>
</tr>
<tr>
<td>Instrument</td>
<td>7.226</td>
<td>7%</td>
</tr>
<tr>
<td>Rater</td>
<td>10.646</td>
<td>11%</td>
</tr>
<tr>
<td>Subject-by-instrument</td>
<td>3.342</td>
<td>3%</td>
</tr>
<tr>
<td>Subject-by-rater</td>
<td>32.870</td>
<td>34%</td>
</tr>
<tr>
<td>Rater-by-instrument</td>
<td>0.512</td>
<td>1%</td>
</tr>
<tr>
<td>Residual</td>
<td>28.865</td>
<td>29%</td>
</tr>
<tr>
<td>Total</td>
<td>98.265</td>
<td></td>
</tr>
<tr>
<td>$\phi$</td>
<td>.30</td>
<td></td>
</tr>
</tbody>
</table>

Note. Negative estimated variance components were set to zero.

A second GT analysis (see Table 7) was employed to assess the variance in specific scores measuring executive functions attributed to rater, instrument, and dimension (represented by the 6 subscales from each instrument; see Table 1 in Appendix A). Of note, this model employed the BRIEF2 (Gioia et al., 2015a) Plan/Organize subscale and the CEFI (Naglieri & Goldstein, 2013a) Planning subscale; it excluded the CEFI Organizing subscale. This GT analysis employed a three-facet model design, with rater, instrument, and dimension (i.e.,
specific executive functioning constructs) specified as the three facets. The dimension facet represented the rating scale subscales hypothesized to measure theoretically related constructs. This model was a random-effects rater (parent and adolescent) by instrument by dimension (subscale scores) design, and all facets were specified as random effects. Variance components were calculated using the subscale scores for the subject, rater, instrument, dimension, and for all interactions. The variance estimate attributed to differences in subscale scores was considered the universe score variance. The magnitude of the estimated variance components was computed, and the absolute error variance was calculated for the variance estimates attributed to the rater, instrument, dimension, all interactions, and to the residual variance. Dependability coefficients were calculated to determine the overall index of dependability across rater, instrument, and dimension for the subscale scores.

Table 7 provides the variance components estimates for subject, rater, instrument, dimension, and all interactions. The object of measurement, variance attributable to individual differences across subjects, only accounted for 6% of variance in subscale scores; the remainder of variance was due to systematic or random error. Results revealed that the largest proportion of variance in subscale scores was attributed to the subject-by-rater interaction; it accounted for 30% of variance. The subject-by-instrument interaction contributed the next largest proportion of variance (13%). The other facets contributed minimal variance in subscale scores. Specifically, the rater and instrument facets each contributed 6% of variance in subscale scores, and the subject-by-dimension interaction contributed 5% of variance. Furthermore, the rater-by-dimension and instrument-by-dimension facets only contributed 1% of variance. The dimension facet, rater-by-instrument interaction, and rater-by-instrument-by-dimension interaction did not contribute any variance in subscale scores. Residual variance was 32%. The dependability
A third GT analysis (see Table 8) was employed to assess the variance in specific scores measuring executive functions attributed to rater, instrument, and dimension. This model was
identical to the prior model (see Table 7) but substituted the CEFI (Naglieri & Goldstein, 2013a) Organization subscale for the CEFI Planning subscale (see Table 8). Results from this GT analysis generally were consistent with the results obtained from the second GT analysis. Specifically, the variance attributable to individual differences across subjects only accounted for 6% of the variance in subscale scores; the remainder of variance was due to systematic or random error. The largest proportion of variance in subscale scores was attributed to the subject-by-rater interaction (31%). The subject-by-instrument interaction contributed the next largest proportion of variance (11%), though the proportion of variance attributed to this interaction is slightly smaller in comparison to the prior model. The other facets contributed negligible variance in subscale scores; the rater facet, instrument facet, and the subject-by-dimension interaction each contributed 6% of variance. Furthermore, the rater-by-instrument and instrument-by-dimension facets only contributed 1% of variance. The dimension facet, rater-by-dimension interaction, and rater-by-instrument-by-dimension interaction did not contribute any variance in subscale scores. Residual variance was 33%. The dependability coefficient was .16, which again, indicates poor dependability of subscale scores across executive functioning instruments, executive functioning domains, and the raters.

Given that the previous models demonstrated that the dimension facet and its interactions contributed negligible variance in subscale scores (i.e., 0% to 6%), evidence suggests that conducting partial models examining specific dimensions is not warranted. However, due to the large proportion of variance attributed to the rater facet across both models (30% to 31% of variance in subscale scores), partial models examining the rater effect were conducted.


Table 8

Variance Component Estimates and Absolute Dependability Coefficients by Subscales (Including the CEFI Organization Subscale)

<table>
<thead>
<tr>
<th>Facet</th>
<th>Subscale Score</th>
<th>Percent of variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>5.966</td>
<td>6%</td>
</tr>
<tr>
<td>Instrument</td>
<td>6.661</td>
<td>6%</td>
</tr>
<tr>
<td>Rater</td>
<td>5.928</td>
<td>6%</td>
</tr>
<tr>
<td>Dimension</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Subject-by-instrument</td>
<td>11.660</td>
<td>11%</td>
</tr>
<tr>
<td>Subject-by-rater</td>
<td>33.269</td>
<td>31%</td>
</tr>
<tr>
<td>Subject-by-dimension</td>
<td>6.664</td>
<td>6%</td>
</tr>
<tr>
<td>Instrument-by-rater</td>
<td>0.566</td>
<td>1%</td>
</tr>
<tr>
<td>Instrument-by-dimension</td>
<td>1.028</td>
<td>1%</td>
</tr>
<tr>
<td>Rater-by-dimension</td>
<td>0.464</td>
<td>0%</td>
</tr>
<tr>
<td>Rater-by-instrument-by-dimension</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Residual</td>
<td>34.994</td>
<td>33%</td>
</tr>
<tr>
<td>Total</td>
<td>107.200</td>
<td></td>
</tr>
<tr>
<td>( \phi )</td>
<td>.16</td>
<td></td>
</tr>
</tbody>
</table>

Note. Negative estimated variance components were set to zero.

**Partial models.** Additional analyses were conducted to better understand the source of sizable error variance components in subscale scores attributed to each rater (see Tables 9, 10, 11, and 12). Four partial models were analyzed, with one of the raters (i.e., parent or adolescent) omitted from each. All partial models employed a two-facet design, with instrument (i.e.,
BRIEF2 and CEFI) and dimension (i.e., represented by the 6 subscales from each instrument; see Table 1 in Appendix A) specified as the two facets. Additionally, the models employed a random-effects instrument by dimension design. Variance components, the magnitude of the estimated variance components, and the absolute error variance were calculated utilizing the subscale scores for the instrument, dimension, all interactions, and to residual variance. The variance estimate attributed to differences in subscale scores was considered the universe score variance. In addition, dependability coefficients were calculated to determine the overall index of dependability across instrument and dimension for the subscale scores.

In the partial models including only parent raters, the variance due to the object of measurement (i.e., subject) increased from 6% in the full models to 40% and 41% in the partial models including the CEFI (Naglieri & Goldstein, 2013a) Planning subscale and the CEFI Organization subscale, respectively (see Tables 9 and 11). Notably, the variance component for the subject facet was larger in the partial models including only parent raters (i.e., 40% to 41%) than in the partial models including only adolescent raters (i.e., 28%); the size of this variance component remained unchanged in the partial models for the adolescent raters (see Tables 10 and 12). The increase in the size of the variance component for the subject facet suggests that the amount of variance attributable to individual differences across adolescents being rated increases as a function of the parent raters. Similarly, in comparison to the full models, the size of the variance components in the partial models for the parent raters reduced for the instrument facet, the instrument-by-dimension interaction, and the residual variance; the subject-by-dimension interaction slightly increased, which likely is due to adolescent raters producing higher mean subscale scores than parent raters (see Table 3). In the partial models for the adolescent raters, the size of the variance component increased for the instrument facet, the subject-by-dimension
interaction, and the residual variance. Notably, across all of the partial models, the subject-by-instrument interaction accounted for 20% to 26% of the variance in subscale scores; this is a significant increase in comparison to the variance attributed to this interaction in the full models (i.e., 3% to 11%). As expected, given the results from the GT analyses in the prior models, the dimension facet and the instrument-by-dimension interaction remained unchanged across all of the partial models (i.e., contributing 0% of variance in subscale scores). Taken together, the results from the partial models indicate that the majority of the subscale score variance in the full models was due to the inclusion of the rater facet, and most notably to the adolescent raters.

Table 9

<table>
<thead>
<tr>
<th>Estimated variance components</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Facet</td>
<td>Subscale Score</td>
<td>Percent of variance</td>
</tr>
<tr>
<td>Subject</td>
<td>40.842</td>
<td>41%</td>
</tr>
<tr>
<td>Instrument</td>
<td>4.439</td>
<td>4%</td>
</tr>
<tr>
<td>Dimension</td>
<td>0.078</td>
<td>0%</td>
</tr>
<tr>
<td>Subject-by-instrument</td>
<td>25.984</td>
<td>26%</td>
</tr>
<tr>
<td>Subject-by-dimension</td>
<td>10.125</td>
<td>10%</td>
</tr>
<tr>
<td>Instrument-by-dimension</td>
<td>0.307</td>
<td>0%</td>
</tr>
<tr>
<td>Residual</td>
<td>18.776</td>
<td>19%</td>
</tr>
<tr>
<td>Total</td>
<td>100.551</td>
<td></td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.69</td>
<td></td>
</tr>
</tbody>
</table>

Note. Negative estimated variance components were set to zero.
Table 10

Variance Component Estimates and Absolute Dependability Coefficients by Subscales for the Adolescent Rater and Including the CEFI Planning Subscale

<table>
<thead>
<tr>
<th>Facet</th>
<th>Subscale Score</th>
<th>Percent of variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>28.587</td>
<td>28%</td>
</tr>
<tr>
<td>Instrument</td>
<td>8.613</td>
<td>9%</td>
</tr>
<tr>
<td>Dimension</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Subject-by-instrument</td>
<td>21.924</td>
<td>22%</td>
</tr>
<tr>
<td>Subject-by-dimension</td>
<td>12.114</td>
<td>12%</td>
</tr>
<tr>
<td>Instrument-by-dimension</td>
<td>1.430</td>
<td>1%</td>
</tr>
<tr>
<td>Residual</td>
<td>28.426</td>
<td>28%</td>
</tr>
<tr>
<td>Total</td>
<td>101.085</td>
<td></td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.59</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Negative estimated variance components were set to zero.
### Table 11

*Variance Component Estimates and Absolute Dependability Coefficients by Subscales for the Parent Rater and Including the CEFI Organization Subscale*

<table>
<thead>
<tr>
<th>Facet</th>
<th>Subscale Score</th>
<th>Percent of variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>39.173</td>
<td>40%</td>
</tr>
<tr>
<td>Instrument</td>
<td>3.986</td>
<td>4%</td>
</tr>
<tr>
<td>Dimension</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Subject-by-instrument</td>
<td>23.846</td>
<td>24%</td>
</tr>
<tr>
<td>Subject-by-dimension</td>
<td>11.513</td>
<td>12%</td>
</tr>
<tr>
<td>Instrument-by-dimension</td>
<td>0.571</td>
<td>1%</td>
</tr>
<tr>
<td>Residual</td>
<td>19.125</td>
<td>19%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>98.214</strong></td>
<td></td>
</tr>
<tr>
<td>$\phi$</td>
<td>$0.69$</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Negative estimated variance components were set to zero.
Table 12

Variance Component Estimates and Absolute Dependability Coefficients by Subscales for the Adolescent Rater and Including the CEFI Organization Subscale

<table>
<thead>
<tr>
<th>Facet</th>
<th>Subscale Score</th>
<th>Percent of variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>28.652</td>
<td>28%</td>
</tr>
<tr>
<td>Instrument</td>
<td>9.97</td>
<td>10%</td>
</tr>
<tr>
<td>Dimension</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Subject-by-instrument</td>
<td>20.408</td>
<td>20%</td>
</tr>
<tr>
<td>Subject-by-dimension</td>
<td>14.224</td>
<td>14%</td>
</tr>
<tr>
<td>Instrument-by-dimension</td>
<td>0.787</td>
<td>1%</td>
</tr>
<tr>
<td>Residual</td>
<td>28.163</td>
<td>28%</td>
</tr>
<tr>
<td>Total</td>
<td>102.204</td>
<td></td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.59</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Negative estimated variance components were set to zero.
Discussion

The current study not only expanded the relatively small literature base examining the psychometric properties of executive function rating scales, but it also addressed the weaknesses in prior rating scale research by providing a nuanced interpretation of the reliability, validity, and generalizability of test scores obtained from parent-adolescent reports on the BRIEF2 (Gioia et al., 2015a) and the CEFI (Naglieri & Goldstein, 2013a) rating scales—two instruments widely used in clinical practice to measure executive function. Historically, CTT has been the only model utilized to examine the sources of error variance from rating scales of executive function; this method presents limitations in fully identifying multiple sources of variance in test score, as CTT analyses generally examine only one source of error variance at a time and only yield a global error term (Briesch et al., 2014). This limitation is highly problematic given that it is well-supported that influences such as source and setting, instrument, and dimension frequently contribute to the variance in rating scales’ test scores (Bergeron et al., 2008; De Los Reyes & Kazdin, 2005; Duhig, et al., 2000; Grietens et al., 2003; Kramer, 2006; Reynolds et al., 2009).

The current study is unique in comparison to prior studies examining the psychometric properties of executive function rating scales because, in addition to conducting classical CTT analyses, this study employed GT analyses in order to simultaneously identify the sources of error that disrupt the dependability of broad and specific scores from executive function rating scales. As such, this study expanded the current literature by (1) identifying the proportion of error variance attributed to the instrument, the rater, the instruments’ targeted dimensions, and to their interactions in global and specific scores of executive function across parent-adolescent reports; (2) evaluating the overall dependability of scores obtained from parent-adolescent reports on rating scales of executive function; and (3) providing insight regarding the validity and reliability
of adolescents’ self-reports on behavioral rating scales, and more specifically, on rating scales of executive function, as research in this area has yielded inconsistent results, and is minimal to say the least (Baker et al., 2007; De Los Reyes & Kazdin, 2005; Grietens, 2004; Kramer, 2006; Karver, 2006; Willard et al., 2016; Youngstrom et al., 2003).

**Dependability of Scores for Executive Function Rating Scales**

Dependability coefficients are integral in understanding the extent to which a single score, and in the current study, the global and specific executive function scores, can be generalized to other scores measuring the same behavioral domain from different facets (e.g., instrument, rater, and dimension; Shavelson & Webb, 1991). Across all of the GT analyses conducted in this study, none of the dependability coefficients (i.e., ranging from .16 to .69) for the executive function rating scales met the typical lower level boundary of .80 for acceptable reliability; in fact, all of the dependability coefficients failed to meet even the much lower, though acceptable, reliability coefficient criterion of .70 or higher used for screening instruments (Salvia, Ysseldyke, & Bolt 2010).

The dependability coefficients in this study were markedly low (i.e., ranging from .16 to .30) and were generally much lower than expected given the corresponding Pearson correlation coefficients evaluating the effect of instrument, rater, and dimension in isolation, as well as the results from reliability analyses of executive function rating scales in prior research. Specifically, in the GT analysis evaluating the collective effects of instrument, rater, and their interactions, the resulting dependability coefficient was .30. Moreover, GT analyses yielded significantly lower dependability coefficients when instrument effects, rater effects, dimension effects, and their interactions were considered collectively; the resulting dependability coefficient was .18 and .16, with the inclusion of the CEFI (Naglieri & Goldstein, 2013a).
Planning subscale and the CEFI Organization subscale, respectively. These dependability coefficients are much lower than the moderate to strong correlations (e.g., ranging from .67 to .95) between the total scores and subscale scores for each instrument in isolation; lower than the moderate correlations between instrument total scores (i.e., ranging from .33 to .61); and lower than the weak to moderate correlations between instrument subscale scores (i.e., ranging from .11 to .65). It is also notable that the resulting dependability coefficients in this study were weaker than the strong to very strong mean internal consistency reliability coefficients produced for each instrument (i.e., coefficients generally ranging from .90 and above for total scores and .80 and above for subscale scores), according to the BRIEF2 Professional Manual (Gioia et al., 2015b) and the CEFI Technical Manual (Naglieri & Goldstein, 2013b).

This pattern of low correspondence between raters’ reports of adolescents’ behavior is broadly consistent with several of the findings in rating scale research (Achenbach et al., 1987; Achenbach 2011; De Los Reyes & Kazdin, 2005; Duhig et al., 2000; Grietens et al., 2003; Youngstrom, Loeber, & Stouthamer-Loeber, 2000). Partial GT analyses evaluating instrument and dimension effects collectively for each rater in isolation resulted in relatively higher dependability coefficients (ranging from .59 to .69 for adolescent and parent raters, respectively). Nonetheless, the results from the GT analyses in this study yielded highly questionable dependability coefficients, which indicates extremely poor dependability of global scores and subscale scores for rating scales of executive function. Moreover, these results emphasize the importance of using extreme caution when generalizing subscale scores of executive function across different instruments, raters, and specific executive functions.

**Instrument effects.** The influence of instrument characteristics on global and specific scores of executive function was assessed via several GT analyses, intercorrelations across
scores for each instrument, and a series of paired samples $t$-tests comparing the total scores and the subscale scores for each instrument. The GT analyses in this study that considered the instrument facet and the facets representing interactions with instrument effects revealed that the instrument component contributed 7% variance in total scores and that the subject-by-instrument interaction (3%) and the rater-by-instrument interaction (1%) contributed negligible variance in total scores. Broadly similar variance component estimates were found for the subscale scores for the instrument component (6%), the subject-by-instrument interaction (11% to 13%), the instrument-by-rater interaction (0% to 1%), and the rater-by-instrument-by-dimension interaction (0%). These values are relatively small and support the hypothesis that the rating scale in isolation would contribute a small proportion of variance in executive function scores. This relatively small percentage of variance may be attributable to the similarity in the instruments format (e.g., pencil and paper administration) and that only a few scores, isolated to adolescent raters, had significantly different mean values (e.g., total scores, and shift and flexibility and working memory subscale scores), according to analyses from paired samples $t$-tests. For example, the adolescents’ ratings for the BRIEF2 (Gioia et al., 2015a) were generally higher than their ratings for the CEFI (Naglieri & Goldstein 2013a), which may, in part, be attributable to instrument characteristics. Interestingly, the variance component estimates for the subject-by-instrument interaction in the partial models accounted for about a quarter of the variance (i.e., 20% to 26%), which is notably larger than the variance attributed to this interaction in the full models (i.e., 3% to 13%). This relatively large percentage of variance suggests that the subjects’ (e.g., adolescents being rated) scores varied as a function of instrument, such that some subjects obtained higher or lower scores on one instrument than on the other instrument. According to the mean scores obtained for each instrument in the current study, evidence suggests that
subjects consistently obtained higher scores on the BRIEF2 than on the CEFI. It is possible that instrument characteristics, such as time period, scale anchors, behavioral descriptions of specific executive functions, and norm blocks used to derive norm-referenced scores for each instrument differentially affected obtained scores.

Pertaining to convergent validity indicated by correlations across instruments’ scores, the results are commensurate with previous findings evaluating the convergent validity of executive function rating scales. Prior studies provide evidence for weak to moderate relations between scores from executive function rating scales and scores from relevant executive function scales from other behavioral rating scales (Gioia et al., 2000; Gioia et al., 2015b; Goulden & Silver, 2009; Jarratt, Riccio, & Siekierski, 2005). Notably, the uncorrected correlations obtained in the current study between the total scores from the BRIEF2 (Gioia et al., 2015a) and the CEFI (Naglieri & Goldstein, 2013a) parent ($r = .61$) and self-report ($r = .59$) rating scales were somewhat lower in comparison to the uncorrected correlations obtained in the only existing study examining the convergent validity between the total scores from the BRIEF (Gioia et al., 2000) and the CEFI parent ($r = .76$) and self-report ($r = .79$) rating scales (Naglieri & Goldstein, 2013b).

It is possible that instrument characteristics such as the time period embedded in the instruments’ directions, the instruments’ different scale anchors for rating behaviors, and the instruments’ behavioral descriptions of specific executive functions contributed to significant differences in instrument scores for the adolescent raters. With regard to time period, the BRIEF2 (Gioia et al., 2015a) directions instruct adolescents to rate the frequency of problem behaviors during the past 6 months, whereas the CEFI (Naglieri & Goldstein, 2013a) directions instruct adolescents to rate the frequency of problem behaviors during the past 4 weeks.
Therefore, it is likely that adolescents had difficulty evaluating their own behaviors within these specified time periods, such that they were more likely to over-report behaviors occurring within the past 6 months and under-report behaviors occurring within the past 4 weeks. However, findings from Lakes and Hoyt (2009) suggest that informant response bias, as opposed to instrument bias, generally accounts for a significant amount of error in scores produced from behavior rating scales because informant ratings require the application of memory and inference, which is highly subjective and prone to error. These findings suggest that the time period confound probably was minimized, as adolescent raters were more likely to remember more recent behaviors across both instruments, as opposed to behaviors that have occurred 4 weeks or 6 months ago.

With regard to the instruments’ different scale anchors for rating behaviors, the BRIEF2 (Gioia et al., 2015a) form contains items that are rated on a 3-point scale (from Never to Often). In contrast, the CEFI (Naglieri & Goldstein, 2013a) form contains items that are rated on a 6-point scale (from Never to Always). This difference in scaling may contribute to the significant differences identified in the adolescents’ total scores and select subscale scores given that the difference in scaling presents more or less subjectivity in ratings. For example, the relatively limited scaling for the BRIEF2 may influence raters to choose extreme values (i.e., Never and Always), while less stringent scaling for the CEFI may influence raters to choose less extreme values, given that raters are presented with 3 additional ratings (e.g., Rarely, Very Often, and Always). Close examination of the BRIEF2 and CEFI self-report forms reveals that the BRIEF2 presents items in absolute statements (e.g., “I have angry outbursts”) and that the CEFI presents items in question form (e.g., “How often… do you find it hard to control your emotions?”). In addition, all of the items included in the BRIEF2 are worded negatively, while some of the items
included in the CEFI are worded positively (i.e., 23). These linguistic characteristics of the instruments may have contributed to the differences in the adolescents’ total scores and selected subscale scores. Furthermore, depending on how carefully adolescents read and responded to items and on the adolescent’s reading level, it is probable that the linguistic characteristics of the instruments are more likely to contribute to the variation in rating scale scores. However, according to the BRIEF2 Professional Manual (Gioia et al., 2015b) and the CEFI Technical Manual (Naglieri & Goldstein, 2013b), both instruments were developed to require the lowest reading level. The BRIEF2 Self-Report form was written to require a fifth grade reading skill level, and the overall readability score for the CEFI Self-Report form indicated an early fifth grade reading level (Gioia et al., 2015b; Naglieri & Goldstein, 2013b).

Additionally, given that the construct of executive functioning in children and adolescents is elusive and inconclusive in the literature, it is possible that the instruments’ descriptions of specific executive functions vary, which ultimately influences raters’ reports of the frequency and severity of the targeted behavior. According to the BRIEF2 Professional Manual (Gioia et al., 2015b), the BRIEF2 (Gioia et al., 2015a) rating scales were developed according to Holmes-Bernstein and Waber’s (1990) conceptualization that the construct of executive functioning is multidimensional and multifactorial and represents an interconnected functional system that includes several distinct, yet related cognitive processes involved in complex behaviors and emotional responses. As such, items included in the BRIEF2 were developed to assess specific cognitive processes associated with the broader constructs of executive functioning such as metacognitive processes and behavior regulation. In contrast, according to the CEFI Technical Manual (Naglieri & Goldstein, 2013b), the CEFI (Naglieri & Goldstein, 2013a) rating scales were developed in congruence with the unidimensional theory of
executive functioning, which posits that executive functioning is a regulated system of cognitive processes involving automatic and voluntary behaviors that produce cognitive, behavioral, and emotional responses (Climie, Cadogan, & Goukon, 2014; Fenwick & McRimmon, 2015; Naglieri & Goldstein, 2013b). These different approaches to conceptualizing and measuring the construct of executive function likely influence the instruments’ behavioral descriptions of executive function.

Lastly, it is likely that the norm blocks used to derive norm-referenced scores for each instrument contributed to differences in subjects’ scores across instruments. Norm blocks, sections within norm tables that contain the normative information for a specific age group, are derived from norm samples; norm samples are employed across measures of psychological constructs in order to compare an individual’s scores to those from a large group of individuals who share similar characteristics (e.g., gender and age). According to Flanagan, Ortiz, Alfonso, and Mascolo (2006), it is important to consider the similarity of individuals’ characteristics as well as the developmental differences attributed to age when developing norm blocks in order to increase the sensitivity of norms and thus, the accuracy of norm-referenced scores. As such, the differences in the norm blocks employed in the BRIEF2 (Gioia et al., 2015a) and the CEFI (Naglieri & Goldstein 2013a) to derive norm-referenced scores likely inflated or deflated scores based on the individuals’ characteristics within the norm block employed.

The BRIEF2 (Gioia et al., 2015a) scores are based on gender and age group, whereas the CEFI (Naglieri & Goldstein 2013a) scores only are based on age group. For each gender, the BRIEF2 Parent form provides norms based on four age groups (i.e., ages 5 to 7, 8 to 10, 11 to 13, and 14 to 18), and the BRIEF2 Self-Report form provides norms based on two age groups (i.e., ages 11 to 14 and 15 to 18). According to the BRIEF2 Professional Manual (Gioia et al.,...
2015b), the norms were separated by gender and age normative groups due to significant differences between scores for gender and age for the BRIEF2 Parent form; the BRIEF2 Self-Report form yielded significant differences for gender and several age groups on most scales. The CEFI Parent form provides combined gender norms based on two age groups (i.e., norm blocks for ages 5 to 11 and 12 to 18), and the CEFI Self-Report form provides combined gender norms based on one age group (i.e., norm block for ages 12 to 18). According to the findings reported in the CEFI Technical Manual (Naglieri & Goldstein, 2013b), significant differences for gender were found for most scales on the CEFI Parent form and only for two scales (i.e., Initiation and Organization) on the CEFI Self-Report form. Overall, fewer executive function deficits were endorsed for females than for males. Combined gender norms were employed in order to reflect the gender differences that exist in the general population (Naglieri & Goldstein, 2013b). Significant differences between scores for ages 5 to 11 and 12 to 18 were found for the CEFI Parent form and no significant age differences (i.e., between ages 12 to 14 and 15 to 18) were found for the CEFI Self-Report form.

According to the mean scores obtained in this study for the total scores and subscale scores across instrument (see Table 3), scores obtained for the BRIEF2 (Gioia et al., 2015a) were consistently higher relative to scores obtained for the CEFI (Naglieri & Goldstein, 2013a); this pattern of scoring is consistent with the norm blocks employed for each instrument as well as to the fact that the majority of the study’s sample included girls. As such, norm block estimates for the BRIEF2 included same-gender norms and norms based on relatively small age ranges (e.g., 2 years for Parent form and 3 years for Self-Report form), which may inflate score estimates for girls because scores obtained for girls and younger ages are not derived relative to boys’ scores or relative to older age groups (Naglieri & Goldstein, 2013a; Sulik et al., 2010). However, norm
block estimates for the CEFI included combined gender norms and norms based on relatively large age ranges (e.g., 6 years for Parent and SELF-Report forms), which may deflate score estimates for girls, as research indicates that girls and older adolescents, on average, exhibit fewer executive function deficits relative to boys and younger children, respectively (Naglieri & Goldstein, 2013a; Sulik et al., 2010).

**Dimension effects.** Surprisingly, the results from the GT analyses did not support the hypothesis that the dimension component would contribute to a large proportion of variance in executive functioning, as the variance components due to the dimension facet and the facets representing interactions with dimension effects generally contributed negligible variance (0% to 1%) in total scores and subscale scores. Larger percentages of variance in total scores (6%) and subscale scores (12% to 16%) were attributable to the subject-by-dimension interaction, which indicates that variance in subscale scores is due to individual differences across executive functioning dimensions, which is expected. Additionally, findings from the intercorrelations within and across instruments, and the paired samples t-tests between scores from the dimensions of interest (e.g., Inhibit and Inhibitory Control, Shift and Flexibility, Emotional Control and Emotion Regulation, Plan/Organize and Planning and Organization, and the Working Memory and Self-Monitor subscales from both instruments) were broadly congruent with the GT analyses.

All of the within instrument intercorrelations between the total score and the subscale scores were strong (i.e., .71 to .89) for the BRIEF2 (Gioia et al., 2015a) and were moderate to very strong (i.e., .67 to .95) for the CEFI (Nalgieri & Goldstein, 2013a). These results are consistent with construct validity studies examining the relations between the total scores and the subscale scores from the BRIEF2 (Gioia et al., 2015a) and the CEFI (Naglieri & Goldstein,
2013a) in isolation. Studies presented in the *BRIEF2 Professional Manual* (Gioia et al., 2015b) and the *CEFI Technical Manual* (Naglieri & Goldstein, 2013b) indicate that both instruments in isolation evidence strong relations between subscale scores for the dimensions of interest and the total score. For example, the intercorrelations for the BRIEF2 were moderate to strong (i.e., .78 to .92) between subscale scores and the total score. Additionally, CFAs for the BRIEF2 indicated strong to very strong (i.e., .83 to .92) relations between subscale scores and three factors. Similarly, scale-level EFAs for the CEFI demonstrated moderate to very strong (i.e., .72 to .94) relations between subscale scores and one factor (i.e., total score). The results from this study support prior research indicating that the construct of executive function likely is unidimensional; the GT analyses demonstrated that the dimension facet—based on the subscale scores proposed to measure similar theoretical constructs—and the interactions with the dimension facet contributed minimal to no variation in total scores and subscale scores across two instruments of executive function.

In contrast, the across instrument intercorrelations between subscale scores for the dimensions of interest ranged from weak to moderate (i.e., .27 to .69), though the majority of the correlations were within the moderate range (i.e., .42 to .69). Moreover, the paired samples *t*-tests revealed relatively few statistically significant differences between the subscale scores contributing to the dimensions, which provides further evidence of construct validity for executive function domains. These findings generally are inconsistent with existing research examining the convergent validity among measures of executive functioning. Although no studies exist that evaluate the construct validity of specific executive function domains across rating scales of executive function, studies examining the convergent validity of executive function rating scales generally report weak to moderate correlations between specific executive
functioning domains and select subscales hypothesized to measure executive functions from
other behavioral rating scales (Achenbach, 1991; Gioia et al., 2015b). Similarly, the majority of
studies examining neuropsychological batteries often yield inconclusive results with regard to
fully substantiating the construct validity of the test (Burns, Nettlebeck, & McPherson, 2009;
Kane, Conway, Miura, & Colflesh, 2009; Keith, Kranzler, & Flannagan, 2001). Overall, the
findings from the current study suggest that examining the validity of specific executive
functions in isolation potentially yields erroneous results given that the results from the GT
analysis indicate that the collective effect of multiple sources of error on scores decreases the
amount of variance attributable to the dimension effect and its interactions, which, in addition to
generally moderate intercorrelations between dimension scores, supports the construct validity of
specific executive function domains.

**Rater effects.** The results from this study yielded negligible to weak inter-rater reliability
coefficients between the total scores and subscale scores from both the BRIEF2 (i.e., .07 to .35;
Gioia et al., 2015a) and the CEFI (i.e., .04 to .42; Naglieri & Goldstein, 2013a). Moreover,
across GT analyses, variance component estimates indicated that the subject-by-rater interaction
contributed the largest proportion of variance in total scores (33%) and subscale scores (30% to
31%); this large value indicates that this interaction had a significant effect on the variation in
total scores and subscale scores measuring executive function. Furthermore, the results from the
partial models indicated that examining each rater in isolation significantly increased the
proportion of variance attributable to the subject facet for both the adolescent raters (from 6% in
the full model to 28% in the partial model) and the parent raters (from 6% in the full model to
40% to 41% in the partial model); these results suggest that most of the subscale score variance
in the full models was due to the inclusion of the rater facet, and particularly to the adolescent
raters. These findings support the hypothesis that the rater component would contribute the largest proportion of variance in executive function scores.

The results from this study are inconsistent with the inter-rater reliability coefficients for the total scores (i.e., .71) and subscale scores of interest (i.e., .46 to .67) reported for parent-adolescent dyads in the *BRIEF2 Professional Manual* (Gioia et al., 2015b); however, they are broadly consistent with the correlations for the subscale scores of interest (i.e., .38 to .67) reported for parent-adolescent dyads in the *CEFI Technical Manual* (Naglieri & Goldstein 2013b). Moreover, the results from this study are commensurate with several studies in rating scale research providing overwhelming evidence that discrepancies often exist between informants’ ratings of adolescent behavior on behavioral rating scales (Achenbach et al., 1987; Achenbach 2011; Bergeron et al., 2008; De Los Reyes & Kazdin, 2005; Duhig et al., 2000; Grietens et al., 2003; Youngstrom, Loeber, & Stouthamer-Loeber, 2000). Furthermore, the findings from this study support that parent-adolescent raters often exhibit a pattern of low correspondence between their reports of adolescent’s behavior on a variety of behavioral rating scales, including rating scales of executive function (Achenbach et al., 1987; Achenbach, 2011; Baron, 2000; De Los Reyes & Kazdin, 2005; Grietens et al., 2003; Walker & D’Amato, 2006; Youngstrom, Loeber, & Stouthamer-Loeber, 2000). In the current study, it appears that the discrepancy in executive function scores was due to the adolescents’ ratings yielding higher mean scores than the parents’ ratings on both instruments, a pattern similar to prior findings in rating scale research, but that has not yet been substantially established for rating scales of executive function (Achenbach, 1991; Baron, 2000; Meredyth, Prout, & Blaha, 2003; Walker & D’Amato, 2006).
Several studies have conceptualized informant discrepancies on ratings of adolescent’s behavior by considering the characteristics of the informants and the characteristics of the adolescents being rated (Achenbach et al., 1987; Duhig et al., 2000; Grills & Ollendick, 2002). Characteristics relative to this study that may have influenced informant discrepancies include:

(a) adolescent age; (b) social desirability; (c) the adolescent’s perceived stress; (d) the type and severity of the behavior problem; and (e) the context in which parents observe behavior. In addition, the Attribution Bias Context (ABC) Model proposed by De Los Reyes and Kazdin (2005) provides a sound framework, grounded in theory and research, in which to further conceptualize the discrepancies that surfaced in the current study between parent-adolescent ratings of adolescent’s executive functioning.

The ABC model posits that discrepancies in informant attributions and perspectives result in informant discrepancies in behavior ratings (De Los Reyes & Kazdin, 2005). Research suggests that adolescents are more likely to attribute the cause of their problem behaviors to external factors (e.g., the context or environment) rather than to internal factors (e.g., their own personal characteristics). The tendency for adolescents to internalize this attribution may be a function of age. Studies indicate that due to adolescents’ developmental and cognitive level, they typically lack awareness of their own cognitive and executive functioning deficits and of the negative affect that these deficits have on their functioning across multiple domains (e.g., home, school, and social; Best & Miller, 2010; De Los Reyes & Kazdin, 2005). This lack of insight likely yields inaccurate ratings of adolescents’ self-reports of daily executive functioning, which contributes to discrepancies between parent-adolescent ratings of everyday executive function. Furthermore, given that executive functioning and self-awareness increase with age, research suggests that adolescents may have increased awareness of their executive function deficits yet
an inaccurate interpretation of how these deficits impact their daily functioning (Best & Miller, 2010; De Los Reyes & Kazdin, 2005; Willard et al., 2016). This increased awareness coupled with adolescents’ difficulties in translating their level of awareness into accurate and valid self-reports likely leads to adolescents falsely over-reporting deficits in executive function (Best & Miller, 2010; Willard et al., 2016). As such, converging evidence supports that adolescents’ lack of awareness of their own behavior, and of the broader impact of their behavior on the environment, likely results in informant discrepancies, especially because parents are more likely to attribute adolescent’s behavior to internal factors (De Los Reyes & Kazdin, 2005).

In addition, rater social desirability potentially contributes to false attributions and perspectives of adolescent’s executive functioning. Research suggests that adolescents are more likely to rate their behavioral deficits, particularly those associated with aggressive behavior and attention problems, as less severe than in reality in an attempt to normalize their behavior and to conform to social norms in order to present themselves favorably (De Los Reyes & Kazdin, 2005; Meredyth, Prout, & Blaha, 2003). Similarly, parents may under-report behavior deficits in order to avoid stigmatizing labels placed on adolescents (Norfolk & Floyd, 2016; Ohan, Visser, Strain, & Allen, 2010). Within the context of the current study, adolescents may have exhibited acquiescent responding and endorsed executive functioning deficits to a higher degree given their awareness that they were participating in a study examining the consistency in parent and adolescent reports of executive functioning and therefore attempted to respond to items in an acquiescent way given the context (i.e., for the study to find “results”). Conversely, parents may have displayed socially desirable responding and endorsed adolescents’ executive functioning deficits to a lower degree for the same reason.
Other factors that contribute to informant discrepancies include parents’ attributions and perceptions of adolescent behavior. Given that the developmental time period of adolescence is characterized by increased autonomy, in which the amount of time adolescents spend with parents typically declines, it is possible that parents do not have adequate exposure to adolescent’s behavior and therefore, are unable to observe deficits in executive functioning with accuracy and validity. Furthermore, deficits in executive function are difficult for parents to directly observe and to interpret the associated functional daily outcomes, as executive function deficits are less salient and less overtly bothersome than externalizing behaviors (Isquith et al., 2013). These barriers increase the occurrence of parental misattributions and misperceptions regarding the frequency and severity of the adolescent’s executive functioning, such that parents are more likely to under-report executive function deficits. Such factors inevitably increase the likelihood of parent-adolescent misperception, which consequently decrease the consistency between parent-adolescent reports of executive function.

It is also noteworthy that the majority of the participants in the study were female. Specifically, 81% of adolescent and 88% of caregiver informants in the study were female, which may account for the variance in scores attributed to the rater facet and its interactions. However, the study’s informant sample closely matched the characteristics of instruments’ informant sample. According to the *BRIEF2 Professional Manual* (Gioia et al., 2015b) and the *CEFI Technical Manual* (Naglieri & Goldstein, 2013b), the majority of informants included in the instruments’ norming samples were mothers, which is consistent with informant data from other rating scales. Furthermore, analyses provided in the *BRIEF2 Professional Manual* indicated no significant differences in scores between informants. On the other hand, the informants included in the norming sample for the BRIEF2 (Gioia et al., 2015a) and CEFI
(Naglieri & Goldstein, 2013a) Self-Report forms were closely matched by gender, though slightly more females were included in the BRIEF2 Self-Report norming sample. The discrepancy between the quantity of adolescent females being rated included in the current study and of those included in the instruments’ norming samples may provide rationale for the relatively low intercorrelations obtained within and across instruments in the current study, given that the current sample was unbalanced with regard to the quantity of adolescent males being rated.

Limitations and Directions for Future Research

The results of this study provide insight into the construct validity and generalizability of scores obtained from executive function rating scales, as well as the utility of adolescents’ self-reports of executive function; however, these results should be interpreted with caution, as they are associated with a number of weaknesses. First, it is possible that some sources of variance in models were overlooked, despite grounding all hypothesized models in theory and research from reputable sources and testing every source of error variance that seemed reasonable for the purpose of the study. Therefore, additional GT analyses would be beneficial in identifying other plausible sources of error across rating scales of executive function, especially given that the literature focusing on the validity of executive function rating scales is remarkably sparse, and generally weak and inconclusive. With regard to overlooked sources of error, it would be beneficial to expand the current study in order to evaluate variance in executive function scores attributable to parent, self-report, and teacher ratings. Examining all three sources of ratings simultaneously would parse apart the error attributable to the rater facet and to all its interactions, which will provide further insight into not only the dependability of executive function scores but also the utility in collecting data from multiple informants.
Additionally, it is possible that the scores produced by the BRIEF2 (Gioia et al., 2015a) and the CEFI (Naglieri & Goldstein, 2015a) are confounded by time. The BRIEF2 directions instruct informants to rate the frequency of problem behaviors during the past 6 months and the CEFI directions instruct informants to rate the frequency of problem behaviors during the past 4 weeks, which introduces the confound of time. It would be beneficial for future research to control for the confound of time via designing an experimental study to randomize the BRIEF2 and CEFI directions in a two-by-two design or by modifying both of the rating scales’ directions to only one of the time periods in order to create consistency in time period for rating behaviors.

Second, this study only included school-age children, ages 12 to 17 years, enrolled in general and special education in a rural, public middle and high school in Florida. Additionally, the majority of the participants in this study were female. According to the demographic data collected by the National Center for Education Statistics (NCES) during the 2016-2017 academic year, 70% of enrolled students are White, 21% of enrolled students are Black, 2% of enrolled students are Asian, 3% of enrolled students are biracial or multiracial, and 3% of enrolled students are Latino/a; the sample in this study closely matches that of the school district. Furthermore, data indicate that 45% percent of students in the district are eligible for free lunch. On state assessments, the majority of students exhibited inadequate proficiency in reading (33%) and in math (36%). Thirty-five percent of students demonstrated below satisfactory proficiency in reading, whereas 24% of students demonstrated this level of proficiency in math. Thirty-one percent of students achieved satisfactory proficiency in math, but only 17% of students obtained this level of proficiency in reading. As such, results obtained in this study may not generalize to populations with characteristics beyond the study’s sample, which poses a limitation to this
study. Moreover, employing a sample from one specific population in GT analyses limits the ability to test the true, best model of error variance.

Third, the sample size for this study was smaller than expected ($N = 42$), though adequate for GT analyses, which do not employ statistical hypothesis testing and therefore, do not yield a probability value. No clear guidelines exist regarding the appropriate sample size for GT analyses, and the literature involving this topic remains murky. Webb, Rowley, and Shavelson (1998) proposed that GT studies should include a minimum of 20 participants and two conditions per facet, though other studies have conducted generalizability analyses with fewer participants and more than two conditions per facet (Briesch et al., 2014). In comparison, findings from a multivariate generalizability study indicated that a sample size of 30 does not provide an adequate estimation of generalizability and dependability coefficients, but that sample sizes of 50 to 300 are adequate for their robust estimation (Atilgan, 2013). By recruiting a total of 42 parent-adolescent dyads and examining three facets, with at least two levels per facet, the current study yielded 1,008 data points for each facet. However, a larger sample size would undoubtedly yield a more precise estimation key statistics.

Lastly, this study employed prorated scores for CEFI rating scales with 6 or fewer missing items, and items were scored as 1 (Never) for BRIEF2 ratings scales with 1 missing item; these scoring procedures may have inflated CEFI scores and restricted BRIEF2 scores. Additionally, the majority of the mean scores for the BRIEF2 Parent and Self-Report forms in this sample were slightly greater than the population mean, which indicates expansion of range. Similarly, the mean scores for the CEFI Self-Report form in this sample were slightly greater than the population mean, suggesting expansion of range; however, the mean scores in this sample for the CEFI Parent form were slightly below the population mean, indicating restriction
of range. Therefore, the results from this study may be confounded by sampling error and warrant cautious interpretation. Notably, this sample only included select subscales from the BRIEF2 and the CEFI. For future studies, it would be beneficial to evaluate the variance for all of the scores obtained from the BRIEF2 and CEFI subscales contributing to the global scores.

**Implications for Practice**

Converging evidence from this study suggests that executive function rating scales yield highly suspect dependability in global and specific scores of executive function. Additionally, this study provides further evidence that the global score for executive function is the most reliable score, as there is little unique variance attributable to subscale scores. Moreover, this study demonstrated that parent-adolescent reports of executive functioning are notably discrepant. Although the dependability of ratings was poor for parents and adolescents, the dependability of adolescent ratings was inferior when compared to parent ratings, which suggests that adolescents seem to be poor reporters of their own executive functioning. Additionally, it appears that the raters’ total scores varied as a function of rater and instrument, such that adolescent raters endorsed more deficits on the BRIEF2 (Gioia et al., 2015a) than on the CEFI (Naglieri & Goldstein, 2013a). Most notably, the results from this study also imply that the construct of executive function remains ambiguous, as raters of behavioral descriptions of executive function appeared to struggle to focus on its central features in any organized manner.

This study sheds light on the construct validity of executive function rating scales, which informs clinical practice in use and interpretation of executive function rating scales. As such, based on the findings from this study, clinicians are advised to rely on the global executive function score for interpreting scores from rating scale of executive function. This score was the most reliable across the CTT analyses and GT analyses in this study, a pattern that also is
consistent with findings evaluating performance-based measures of executive function, especially for adolescents (Eslinger, 1996; Jensen, 1998; Reynolds, Alexander et al., 2009; Reynolds, Graybill et al., 2009). Notably, the fact that the dimension facet and its interactions contributed little to no variance in scores challenges the uniqueness (and perhaps utility) of executive function rating scales. Additionally, clinicians should be mindful that the evidence provided in technical manuals that examines the consistency within and across instrument and rater only evaluates the variance attributed to these facets in isolation and neglect to simultaneously consider the effect of multiple sources of variance on obtained executive function scores. Furthermore, clinicians should attend to the norm blocks employed to derive norm-referenced scores for the adolescent being rated and be particularly cautious when interpreting scores across behavioral rating scales, as the norm blocks employed may inflate or deflate scores based on the characteristics included in the instruments’ norm blocks. Such discrepancies in instruments’ norm blocks alter the interpretation of the resulting scores and thus, the accuracy of population-based inferences drawn from test scores.

Based on these results, clinicians should use extreme caution when using executive function rating scales to identify deficits in executive function. Clinicians are highly discouraged to use parent-adolescent reports on executive functioning rating scales as the sole measure of executive function because deficits in executive function may or may not be detected, and more notably, the frequency and severity of the deficits may be over- or under-represented. Instead, assessment administration should include multisource and multimethod approaches to measuring executive functioning, externalizing behaviors, and internalizing behaviors in adolescents. Furthermore, if possible, clinicians should strive to administer all three executive functioning forms (i.e., parent, teacher, self-report), in order to increase the reliability of the scores. It is also
important for clinicians to consider the developmental, cognitive, and reading level of the raters in order to ensure that raters fully understand the rating scales’ instructions and item questions. If necessary, as specified in both the BRIEF2 (Gioia et al., 2015a) and the CEFI (Naglieri & Goldstein, 2013a) technical manuals, clinicians should verbally read the instructions and items aloud to the rater if the raters’ reading level and level of cognitive functioning is questionable.
References


Appendix A

Table 1

*Cross Reference Domains for the Behavior Rating Scale of Executive Function, Second Edition and the Comprehensive Executive Function Inventory*

<table>
<thead>
<tr>
<th>Behavior Rating Scale of Executive Function, Second Edition</th>
<th>Comprehensive Executive Function Inventory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Executive Composite</td>
<td>Full Scale</td>
</tr>
<tr>
<td>Inhibit</td>
<td>Inhibitory Control</td>
</tr>
<tr>
<td>Shift</td>
<td>Flexibility</td>
</tr>
<tr>
<td>Self-Monitor</td>
<td>Self-Monitoring</td>
</tr>
<tr>
<td>Emotional Control</td>
<td>Emotion Regulation</td>
</tr>
<tr>
<td>Plan/Organize</td>
<td>Planning</td>
</tr>
<tr>
<td>Working Memory</td>
<td>Working Memory</td>
</tr>
</tbody>
</table>

Appendix B

105
LETTER OF INVITATION

Dear Parent or Guardian:

Your child’s school has agreed to allow parents and students from the School District to participate in a study to better understand the consistency between parent’s and children’s ratings of children’s every day attention, self-control, planning, and organizing behaviors. **We are asking your help with this study.**

If you agree to participate in our study and to allow your child to participate in our study, you will complete two forms focusing on their behaviors and your child will complete two forms during school hours. As part of this study, you and your child’s responses on the questionnaires will be shared with only the primary investigator and other researchers for data purposes. You and your child’s names and responses on the forms will be kept confidential within the limits allowed by law, and participation is voluntary. You or your child will not be placed in any harm by taking part in our study. Participation in this study is completely voluntary and you are free to withdraw you and your child from the study at any time without giving a reason and without penalty. You can also have you and your child’s information removed from the research record or destroyed.

Because we recognize and appreciate the notable time and effort required of you, your child, and your child’s school, we want to thank you all for participating in this study. **We want to thank participating parents by giving a hand-out of evidence-based study habits that support children’s academic success as well as the option to enter your name into a raffle to win a $10 Amazon gift card.** Upon completion of you and your child’s participation in the study, the hand-out will be sent home with the child and your name will be entered into a raffle in which you have the chance to win a $10 Amazon gift card, if you choose to participate in the raffle. There will be 15 Amazon gift cards and 50 participants in the study.

If you are willing to allow your child to participate in our study, please complete the following steps:

1. **Carefully read** the *Informed Consent Form*, which explains the details of the study.

2. **Sign** the *Informed Consent Form* to indicate that you give your child permission to participate in this study, and keep a copy for your own records.

3. **Return** the signed *Informed Consent Form* to your child’s guidance counselor or teacher.

If you would like more information about the study before agreeing for you and your child to participate, please contact me in the Psychology Department at The University of Memphis at (901) 219-1060. We thank you for your consideration in participating in the study.

Sincerely,
Leah Singh, M.S.
Principal Investigator
The University of Memphis
Appendix C

Consent to Participate in a Research Study

Understanding Parent-Youth Ratings of Executive Functioning

WHY ARE YOU BEING INVITED TO TAKE PART IN THIS RESEARCH?
You are being invited to take part in a research study about understanding parent’s and child’s ratings of executive functioning across two rating forms. You are being invited to take part in this research study because you are a parent of a child enrolled in the school. If you and your child volunteer to take part in this study, you will be one of about 50 parents and children to do so.

WHO IS DOING THE STUDY?
The person in charge of this study is Leah Singh, M.S. of The University of Memphis Department of Psychology. She is being guided in this research by Dr. Randy Floyd. There may be other people on the research team assisting at different times during the study.

WHAT IS THE PURPOSE OF THIS STUDY?
The purpose of the study is to investigate parent and child ratings of executive functioning skills across two rating forms. Executive functioning skills refer to everyday attention, self-control, planning, and organizing behaviors. Both forms are widely used and have already been evaluated based on assessment of many thousands of children and adults in the United States. We hope to learn about the similarities and differences between ratings from parents and children, the different rating forms, and the different areas of executive functioning.

ARE THERE REASONS WHY YOU SHOULD NOT TAKE PART IN THIS STUDY?
Of children enrolled in the School District, those over the age of 18 will not be allowed to participate in the study.

WHERE IS THE STUDY GOING TO TAKE PLACE AND HOW LONG WILL IT LAST?
I will give or mail you the rating forms you are to complete and ask you to return them to me during the academic year (January-May) of this year. It is expected that it will take approximately 30 minutes for you to read through the instructions and complete the forms.

For your child, completion of the forms will be offered in an individual or group format at your child’s school during school hours (at an agreed upon time by school administrators and teacher) of this academic year. It is expected that it will take approximately 30 minutes for your child to complete the forms.

WHAT WILL YOU BE ASKED TO DO?
As part of this study, you and your child will complete two rating forms in which the frequency and severity of your child’s everyday attention, self-control, planning, and organizing behaviors are rated on a 3-point and 6-point scale.
WHAT ARE THE POSSIBLE RISKS AND DISCOMFORTS?
To the best of our knowledge, the things you will be doing as part of the study have no more risk of harm than you would experience in everyday life, though you may experience a previously unknown risk or side effect.

WILL YOU BENEFIT FROM TAKING PART IN THIS STUDY?
There is no guarantee that you will benefit from taking part in this study. Your willingness to take part in the study, however, may help society as a whole better understand this research topic.

DO YOU HAVE TO TAKE PART IN THE STUDY?
Participation in this study is completely voluntary and you are free to withdraw participation from the study at any time. You will not lose any benefits or rights you would normally have if you choose not to volunteer. You can stop at any time during the study and still keep the benefits and rights you had before volunteering.

IF YOU DON’T WANT TO TAKE PART IN THE STUDY, ARE THERE OTHER CHOICES?
If you do not want to be in the study, there are no other choices except not to take part in the study.

WHAT WILL IT COST YOU TO PARTICIPATE?
There are no costs associated with taking part in the study.

WILL YOU RECEIVE ANY REWARDS FOR TAKING PART IN THIS STUDY?
Upon completion of you and your child’s participation in the study, parents will be given a hand-out of evidence-based study habits that support children’s academic success and the option to enter your name into a raffle to win a $10 Amazon gift card. There will be 15 Amazon gift cards and 50 participants in the study.

WHO WILL SEE THE INFORMATION THAT YOU GIVE?
We will make every effort to keep private all research records that identify you to the extent allowed by law. Your information will be combined with information from other people taking part in the study. When we write about the study to share it with other researchers, we will write about the combined information we have gathered. You will not be personally identified in these written materials. We may publish the results of this study; however, we will keep your name and other identifying information private.

CAN YOUR TAKING PART IN THE STUDY END EARLY?
If you decide to take part in the study you still have the right to decide at any time that you no longer want to continue. If you decide not to take part in this study, your decision will have no effect on the quality of care or services you or your child receives from the school. Further, there are no consequences to withdrawing. If you wish to withdraw, please contact the principal investigator, Leah Singh, ljsingh@memphis.edu, (901) 219-1060.
The individuals conducting the study may need to withdraw you and your child from the study. This may occur if you or your child is not able to follow the study’s directions and if your being in the study is more risk than benefit to you or your child.

WHAT IF YOU HAVE QUESTIONS, SUGGESTIONS, CONCERNS, OR COMPLAINTS?
Before you decide whether to accept this invitation to take part in the study, please ask any questions that might come to mind now. Later, if you have questions, suggestions, concerns, or complaints about the study, you can contact the investigator, Leah Singh, ljsingh@memphis.edu, (901) 219-1060, or Dr. Randy Floyd, rgfloyd@memphis.edu, 901-678-4846. If you have any questions about your rights as a volunteer in this research, please contact the Institutional Review Board staff at The University of Memphis at 901-678-3074.

Please complete the information below to consent you and your child to participate in the study. Return one copy of this form, and keep the second copy for your own records.

Thank you for your time and effort.

I have read the information in the consent form and understand my rights and my child’s rights as a research participant. I understand that I may contact the investigators to answer questions before I or my child participates. Refusal to participate will involve no penalty or loss of benefits to which I am otherwise entitled.

☐ My child and I do want to participate.

☐ My child and I do not want to participate.

_________________________________________  ______________
Signature of parent or guardian  Date

_________________________________________
Printed name of parent or guardian

_________________________________________  ______________
Leah Singh, M.S.  Date
Lead Investigator
Appendix D

DEMOGRAPHIC INFORMATION FORM

If you agree to participate in this study, please complete the demographic questions below about you and your child.

Your telephone # (Used only to provide information about distribution and return of rating form)

Parent Information
What is your relationship to the child? □ Mother  □ Father  □ Other (please specify) ______
What is your age in years? __________
What is your racial background? Please check only one.
  □ African American/Black     □ White/Caucasian     □ Asian/Pacific Islander
  □ Native American/American Indian □ Arab American □ Biracial or Multiracial
  □ Other (please specify) ____________________________
Are you of Hispanic descent?
  □ Yes  □ No
Is English your primary language? □ Yes □ No
  If no, what is your primary language? ____________________________
What is the highest degree you have completed? Please check only one.
  □ Doctorate/Specialist □ Master’s □ Bachelor’s
  □ High School Diploma □ GED □ Other
Child Information

What is your child’s name? ______________________________________

First name          Last name

What is your child’s date of birth? ________________

Month/ Day/ Year

Which gender best describes your child? □ Female □ Male □ Other

What school does your child attend? ________________________________

Has your child been diagnosed with an emotional (e.g., anxiety, depression, etc.) or behavioral (e.g., attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorder (ASD), etc.) disorder?

□ Yes □ No

If yes, what emotional or behavioral disorder(s)?

____________________________________________

Does your child receive special education services (e.g., gifted program, 504 Plan, Individual Education Plan, speech or language therapy at school, etc.)?

□ Yes □ No

If yes, what services does your child receive?

____________________________________________

Is your child of Hispanic descent? □ Yes □ No

Is English your child’s primary language? □ Yes □ No

If no, what is your child’s primary language? __________________________

Which racial background best describes your child?

□ African American/Black □ White/Caucasian □ Asian/Pacific Islander

□ Native American/American Indian □ Arab American □ Biracial or Multiracial

□ Other (please specify) ________________________________
Appendix E

ASSENT FORM

Understanding Parent-Youth Ratings of Executive Functioning

You are invited to be in a research study being done by Leah Singh from The University of Memphis. You are invited because you are enrolled in the School District.

If you agree to be in the study, you will be asked to complete two rating forms during school. You will complete the forms by yourself in an individual or group setting. It will probably take you about 30 minutes or less to complete both forms.

Your family will know that you are in the study. If anyone else is given information about you, they will not know your name. A number or initials will be used instead of your name.

If something makes you feel bad while you are in the study, please tell Leah. If you decide at any time you do not want to finish the study, you may stop whenever you want.

You can ask Leah questions any time about anything in this study. You can also ask your parent any questions you might have about this study.

Signing this paper means that you have read this or had it read to you, and that you want to be in the study. If you do not want to be in the study, do not sign the paper. Being in the study is up to you, and no one will be mad if you do not sign this paper or even if you change your mind later. You agree that you have been told about this study and why it is being done and what to do.

_________________________________________  ______________
Signature of Person Agreeing to be in the Study  Date
Appendix F

Dear Parent:

Thank you for agreeing to participate in the study, *Understanding Parent-Youth Ratings of Executive Functioning*. This study will help us understand the consistency between parent’s and children’s ratings of children’s every day attention, self-control, planning, and organizing behaviors. Please read the following instructions for completing the two rating scales.

**Instructions for completing the BRIEF2 Parent Form**
You will read several statements related to children’s every day attention, self-control, planning, and organizing behaviors. Please think about your child’s behavior over the past 6 months when responding to these questions. Please read each question carefully and rate your child’s behavior on a 3-point scale (*Never (1)* to *Often (3)*). Please try to answer all of the questions. If you make a mistake, just cross it out and circle the correct rating. If you are not certain of an answer, please try to respond the best you can.

Instructions can also be found on the front of the BRIEF2 parent form.

**Instructions for completing the CEFI Parent Form**
You will read several statements related to children’s every day attention, self-control, planning, and organizing behaviors. Please think about your child’s behavior over the past 4 weeks when responding to these questions. Please read each question carefully and rate your child’s behavior on a 6-point scale (*Never (1)* to *Always (3)*). Please try to answer all of the questions. If you make a mistake, just cross it out and circle the correct rating. If you are not certain of an answer, please try to respond the best you can.

Instructions can also be found on the front of the CEFI Parent form.

If you have any questions, please contact me at (901) 219-1060.

Thank you for your time and participation in this study.

Sincerely,
Leah Singh, M.S.
Principal Investigator
The University of Memphis