Title

The Unity and Diversity of Executive Functions: A Systematic Review and Re-Analysis of Latent Variable Studies

Journal

Psychological Bulletin, 144(11)

ISSN

0033-2909

Authors

Karr, Justin E
Areshenkoff, Corson N
Rast, Philippe
et al.

Publication Date

2018-11-01

DOI

10.1037/bul0000160

Peer reviewed

HHS Public Access

Author manuscript

Psychol Bull. Author manuscript; available in PMC 2019 November 01.

Published in final edited form as:

Psychol Bull. 2018 November ; 144(11): 1147–1185. doi:10.1037/bul0000160.

The Unity and Diversity of Executive Functions: A Systematic Review and Re-Analysis of Latent Variable Studies

Justin E. Karr,
University of Victoria

Corson N. Areshenkoff,
University of Victoria

Philippe Rast,
University of California, Davis

Scott M. Hofer,
University of Victoria and Oregon Health & Science University

Grant L. Iverson, and
Harvard Medical School, Spaulding Rehabilitation Hospital, and Home Base, A Red Sox Foundation and Massachusetts General Hospital Program

Mauricio A. Garcia-Barrera
University of Victoria

Abstract

Confirmatory factor analysis (CFA) has been frequently applied to executive function measurement since first used to identify a three-factor model of inhibition, updating, and shifting; however, subsequent CFAs have supported inconsistent models across the lifespan, ranging from unidimensional to nested-factor models (i.e., bifactor without inhibition). This systematic review summarized CFAs on performance-based tests of executive functions and reanalyzed summary data to identify best-fitting models. Eligible CFAs involved 46 samples (N=9,756). The most frequently accepted models varied by age (i.e., preschool=one/two-factor; school-age=three-factor; adolescent/adult=three/nested-factor; older adult=two/three-factor), and most often included updating/working memory, inhibition, and shifting factors. A bootstrap re-analysis simulated 5,000 samples from 21 correlation matrices (11 child/adolescent; 10 adult) from studies including the three most common factors, fitting seven competing models. Model results were summarized as the mean percent accepted (i.e., average rate at which models converged and met fit thresholds: CFI ≥.90/RMSEA ≤.08) and mean percent selected (i.e., average rate at which a model showed superior fit to other models: ΔCFI ≥.005/.010/ΔRMSEA ≤−.010/−.015). No model consistently

Correspondence concerning this article via post should be addressed to Mauricio A. Garcia-Barrera, Department of Psychology, University of Victoria, P.O. Box 1700 STN CSC, Victoria, British Columbia, Canada V8W 2Y2. Electronic correspondence should be addressed to Justin E. Karr (jkarr@uvic.ca).
Justin E. Karr, Department of Psychology, University of Victoria; Corson N. Areshenkoff, Department of Psychology, University of Victoria; Philippe Rast, Department of Psychology, University of California, Davis; Scott M. Hofer, Department of Psychology, University of Victoria, and Department of Neurology, Oregon Health & Science University; Grant L. Iverson, Department of Physical Medicine and Rehabilitation, Harvard Medical School, Spaulding Rehabilitation Hospital, and Home Base, A Red Sox Foundation and Massachusetts General Hospital Program; and Mauricio A. Garcia-Barrera, Department of Psychology, University of Victoria.

converged and met fit criteria in all samples. Among adult samples, the nested-factor was accepted (41–42%) and selected (8–30%) most often. Among child/adolescent samples, the unidimensional model was accepted (32–36%) and selected (21–53%) most often, with some support for two-factor models without a differentiated shifting factor. Results show some evidence for greater unidimensionality of executive function among child/adolescent samples and both unity and diversity among adult samples. However, low rates of model acceptance/selection suggest possible bias towards the publication of well-fitting, but potentially non-replicable models with underpowered samples.

Keywords

executive function; cognitive control; confirmatory factor analysis; latent variable analysis; systematic review

In the past decade, executive functions have garnered a significant amount of clinical and research attention in regard to their definition and measurement (Barkley, 2012; Chan, Shum, Toulopoulou, & Chen, 2008; Jurado & Rosselli, 2007; Pickens, Ostwald, Murphy-Pace, & Bergstrom, 2010). There has also been considerable interest in their predictive validity for clinical and functional outcomes (e.g., childhood problem behaviors; Espy et al., 2011; instrumental activities of daily living; Bell-McGinty, Podell, Franzen, Baird, & Williams, 2002; Cahn-Weiner, Boyle, & Malloy, 2002). Throughout the history of neuropsychology, executive functions have received diverse definitions. Before the term ‘executive functions’ debuted in the neuropsychological literature (Lezak, 1982), researchers had linked the term ‘executive’ with both frontal lobe functioning (Pribram, 1973) and control over lower-level cognitive abilities (Baddeley & Hitch, 1974). However, despite a large body of research on executive functions, the field lacks both a universal definition and an agreed upon form of measurement (Barkley, 2012; Baggetta & Alexander, 2016).

Early models of executive functions detailed a ‘central executive’ that managed lower-level cognitive processes in the context of working memory (Baddeley & Hitch, 1974), while other researchers extended this concept to a system of conscious control over attention (i.e., the Supervisory Attentional System [SAS]; Norman & Shallice, 1986). Based on clinical conceptualizations of frontal processes (Luria, 1966), the functions of the SAS were also attributed to the frontal lobes. These early researchers painted a relatively unitary picture of frontal functioning and executive functions – although they did not yet use this term – where a localized neural substrate underlies a single control function. However, successive definitions of executive functions have illustrated the diversity of abilities falling under this umbrella term (Barkley, 2012; Baggetta & Alexander, 2016); and, further, an established body of neuropsychological research has implicated multiple brain regions that interact with the frontal lobes (e.g., parietal lobes, cerebellum) in the expression of executive functions (Alvarez & Emory, 2006; Collette, Hogge, Salmon, & Van der Linden, 2006; Keren-Happuch, Chen, Ho, & Desmond, 2014).

Prior to unitary models of higher-order cognition, clinicians commonly evaluated many of the abilities now considered executive functions (e.g., planning, self-regulation, fluency) long before scholars clustered these abilities into a common construct (Lezak, 1976). The

debate between the unity and diversity of frontal functioning (Teuber, 1972) and executive functions (Miyake et al., 2000) has continued for decades, although early definitions for executive functions (e.g., Lezak, 1983; Welsh & Pennington, 1988), and nearly all definitions that followed (Barkley, 2012; Baggetta & Alexander, 2016; Jurado & Rosselli, 2007), have described the construct as multidimensional. The earliest definition of executive functions described the construct as having "four components" (Lezak, 1983, p. 507), with later descriptions defining executive functions as an "umbrella term" (Chan et al., 2008, p. 201) for a family of "poorly defined" (Burgess, 2004, p. 79), "meta-cognitive" (Oosterlaan, Scheres, & Sergeant, 2005, p. 69), or "cognitive control" (Friedman et al., 2007, p. 893) processes "used in self-regulation" (Barkley, 2001, p. 5).

Roughly 20 years ago, researchers had already proposed some 33 definitions for executive functions (Eslinger, 1996). The labels and tests for executive functions have been so diverse within the published research that one recent literature review identified 68 subcomponents of executive function, reduced to 18 sub-components following an analysis that removed semantic and psychometric overlap between terms (Packwood, Hodgetts, & Tremblay, 2011). The authors of this review reported that the large number of executive functions posited by various researchers lacked parsimony. In turn, despite years of research on diverse executive functions, the exact number of constructs rightfully labeled executive functions remains largely unknown.

Understanding the number of executive functions supported by the neuropsychological literature first requires an understanding of their measurement. The traditional measurement of executive functions in both research and clinical practice has relied largely on the use of single tests (Baggetta & Alexander, 2016; Chan et al., 2008; Rabin, Barr, & Burton, 2005; Rabin, Paolillo, & Barr, 2016). Tests purported to measure executive functions have varied significantly across studies, with task characteristics sometimes having a greater effect on test performances than the personal and diagnostic features of participants (e.g., age, gender, nature of reading difficulties; Booth, Boyle, & Kelly, 2010). With the heterogeneity of available tests of executive functions, researchers likely inferred that the many tests used to measure executive functions did not all necessarily measure the same unitary construct; however, this inference has resulted in the over-naming of task-specific behaviors as separable executive sub-components (Packwood et al., 2011). This approach ignores the high interrelatedness between both neuropsychological tests and the terms used to describe their outcomes.

Latent Variable Research on Executive Functions

A rich history of published research has explored the correlations between tests of executive functions using a factor analytic approach (Royall et al., 2002). The first factor analyses on executive functions used an exploratory approach that did not impose any hypothesized correlational structure on the battery of tests. The first appearance of an executive function measure in a factor analysis observed the Stroop test loading on a factor involved in the cognitive control over attention (Barroso, 1983). Sequential studies found a heterogeneous number of factors, ranging from a minimum of one factor (e.g., Deckel & Hesselbrock, 1996; Della Sala, Gray, Spinnler, & Trivelli, 1998) to as many as six factors (Testa, Bennett,

& Ponsford, 2012). In multiple contexts, the outcomes of many tasks measuring executive functions loaded together on task-specific factors rather than loading onto common factors composed of indicators from multiple tests (e.g., Cirino, Chapieski, & Massman, 2000; Grodzinsky & Diamond, 1992; Levin et al., 1996; Latzman & Markon, 2010). These findings suggest that the indicators included in these exploratory analyses correlated based on common method variance rather than underlying executive constructs (Barkley, 2012). These task-specific factors may derive largely from the statistical limitations of an exploratory approach, where the relationships between tasks lack a hypothesized structure and potentially group together due to non-executive abilities that also contribute to task performance (Hughes & Graham, 2002).

Many of the tasks employed to measure executive functions have an underlying multidimensional structure (e.g., the Wisconsin Card Sorting Test, Greve et al., 2005; the Trail Making Test, Sanchez-Cubillo et al., 2009), with many different cognitive abilities interacting to explain a given performance (Duggan & Garcia-Barrera, 2015). Executive function tests have a reputation for task impurity, whereby many non-executive abilities explain performances on tests purported to measure executive functions (Burgess, 1997; Miyake & Friedman, 2012; Phillips, 1997). To account for task impurity, a seminal article in the research on executive functions (i.e., Miyake et al., 2000) used a confirmatory factor analysis to assess the relationship between interrelated manifest variables commonly used in cognitive research as measures of three executive functions: the “shifting of mental sets, monitoring and updating of working memory representations, and inhibition of prepotent responses” (p. 50). These researchers constructed a battery of diverse tasks that tapped into three established executive functions, selected based on a rich history of research. They assigned these tasks to hypothesized factors based on their common construct variance and found that a three-factor model best fit the data. In turn, they demonstrated the promise of confirmatory factor analysis at providing purer estimates of executive functions, not contaminated by non-executive method variance. Following this approach, updating, inhibition, and shifting have all received further support through a series of subsequent empirical studies reporting similar three-factor solutions from confirmatory factor models of cognitive tasks (e.g., Friedman et al., 2006, 2008; Lehto, Juujärvi, Kooistra, & Pulkkinen, 2003; Vaughan & Giovanello, 2010).

The published research on measurement models for executive functions has burgeoned in the new millennium (Willoughby, Holochwost, Blanton, & Blair, 2014). The solutions from confirmatory factor analyses accepted by past researchers have varied significantly in terms of the number of factors identified, ranging from a single factor during the preschool and school years (e.g., Brydges, Reid, Fox, & Anderson, 2012; Hughes, Ensor, Wilson, & Graham, 2010; Wiebe, Espy, & Charak, 2008) and older adulthood (e.g., de Frias, Dixon, & Strauss, 2006; Ettenhofer, Hambrick, & Abeles, 2006) to as many as five during young adulthood (i.e., Fournier-Vicente, Larigauderie, & Gaonac’h, 2008). Research on the latent structure of executive function spans all stages of life, but a substantial focus of this research has surrounded the early development of higher-order cognitive abilities (Garon, Bryson, & Smith, 2008; Müller & Kerns, 2015), and a smaller amount of previous work has discussed their development beyond the foundational years of life and into adolescence (Best & Miller, 2010; Best, Miller, & Jones, 2009). Much attention has been given to the differentiation of

executive functions over the course of development (Bardikoff & Sabbagh, 2017), often using a latent variable approach to examine whether factor models support unitary or multidimensional solutions at different ages (e.g., Brydges, Fox, Reid, & Anderson, 2014). Many one-factor solutions have arisen from studies on early executive function development (e.g. Wiebe et al., 2008, 2011; Willoughby, Blair, Wirth, Greenberg, & The Family Life Project Investigators, 2012a), but researchers have criticized the methodology used among young children, where some executive function constructs are rarely evaluated (e.g., shifting) and interpreted as absent, even though they have not been empirically measured by the researchers (Bardikoff & Sabbagh, 2017). Nonetheless, there is evidence for a gradual differentiation of executive function abilities, beginning even prior to the preschool years (Best & Miller, 2010; Garon et al., 2008), where executive functions theoretically transition from a single function to a set of diverse, interactive processes, as many studies on school-aged children, adolescents, and adults found multidimensional solutions of correlated factors (Friedman, Miyake, Robinson, & Hewitt, 2011; Miyake et al., 2000; Lehto et al. 2003).

In terms of cognitive development, the idea of differentiation is not specific to executive functions (Garrett, 1946; Werner, 1957); however, considering the rich empirical research on executive functions in early life, it has gained ground in explaining the changes that occur in the structure of executive functions over the course of development. Some recent interpretations of the executive function literature (Bardikoff & Sabbagh, 2017; Müller & Kerns, 2015) have recruited the interactive specialization framework to explain this differentiation, where cortical areas are functionally non-specific early in life, but over the course of development, become increasingly specialized through activation, interactions, and experience (Johnson, 2000, 2011). Development and organization of basic structural and functional neural networks from birth onwards support greater systems-level integration later in development, particularly within networks that are specialized in executive processing (Luna, Marek, Larsen, Tervo-Clemmens, & Chahal, 2015). Several reviews on the development of executive functions have focused specifically on the neurodevelopment of the three constructs included in the first measurement model reported for executive functions (i.e., inhibition, updating, and shifting; Miyake et al., 2000); however, the factors included in this model do not necessarily represent an exhaustive list of empirically supported executive functions (Jurado & Rosselli, 2007) and, notably, Miyake and colleagues (2000) never described them as such. The terms most commonly used to label executive functions include planning, working memory, fluency, inhibition, and set-shifting (Packwood et al., 2011); however, these terms simply present most frequently in the literature.

The discussion of how many executive functions exist implies that the many abilities labeled “executive” represent separable cognitive capacities; however, each factor does not necessarily represent an orthogonal construct, considering the medium to large correlations often observed between the latent variables of different functions (e.g., .63 to .65, Lehto et al., 2003; .42 to .63, Miyake et al., 2000; .68 to .81, Vaughan & Giovanello, 2010). Working memory capacity and vocabulary both significantly predict outcomes on fluency tasks (Unsworth, Spillers, & Brewer, 2011) and fluency may represent a confluence of working memory interacting with the lexicon (Shao, Janse, Visser, & Meyer, 2014). Similarly, planning represents a higher-order construct, with updating, shifting, and inhibition potentially operating in a collaborative fashion to explain performances on planning-related

tasks (Miyake & Friedman, 2012). The exact relationship between updating, shifting, and inhibition is still not defined, and more recent studies have found that the majority of variance in these three executive functions may be explained by a common higher-order dimension (e.g., Fleming, Heintzelman, & Bartholow, 2016; Friedman et al., 2008; Ito et al., 2015).

Considering the conceptual and empirical overlap between updating, shifting, and inhibition, researchers have begun re-evaluating the shared variance between the constructs through an alternative measurement model (e.g., Friedman et al., 2008, 2011, 2016; Friedman, Corley, Hewitt, & Wright, 2009). Using a nested factor model in repeated analyses of the same dataset, Friedman and colleagues (2008, 2009, 2011, 2016) had all indicators load on a general factor and indicators for updating and shifting co-load on factors specific to those constructs. Because the general factor fully explained the variance in inhibition, the researchers did not include it as a specific factor, with its indicators loading only on the general factor. This model represents an incomplete bifactor model (Chen, West, & Sousa, 2006) and demonstrates a substantial amount of shared variance between indicators across factors in a multidimensional test battery. These findings emphasize the need to consider both general and specific dimensions when explaining performances on test batteries evaluating executive functions.

Aims of the Systematic Review and Re-Analysis

Considering the recent conclusions of Miyake and Friedman (2012) and the many published confirmatory factor analyses supporting multidimensional solutions using performance-based tests (Willoughby et al., 2014), the latent variable research on executive functions has reached a point of requiring both knowledge synthesis and a re-evaluation of previously supported factor solutions. Foremost, the published literature on executive function measurement models has never been comprehensively summarized, and a systematic review would identify the factor models with the most empirical support. Further, few researchers aside from Friedman and colleagues (2008, 2009, 2011, 2016) have evaluated the presence of a common executive function dimension through the nested factor modeling approach described earlier (e.g., Fleming et al., 2016; Garza et al., 2014; Ito et al., 2015; Kramer et al., 2014), but all of these researchers have found a robust general factor. In turn, those researchers not exploring a general dimension potentially over-estimate the diversity of executive function factors. A re-analysis of previous findings would provide a basis to evaluate whether a nested factor model offers superior statistical fit to a multidimensional solution.

The term executive function has become increasingly common within academic literature over the last decade (Willoughby et al., 2014), along with extensive citations of latent variable research (e.g., Miyake et al., 2000) to rationalize the measurement of specific constructs in various research designs. Considering the increased scholarly focus on executive functions, a close assessment of which factor models and constructs have the most empirical evidence will guide researchers when developing their own studies, ensuring their measures target constructs supported by previous scientific inquiry. As well, considering the inferences that have been drawn about the differentiation of executive functions over the

lifespan (Bardikoff & Sabbagh, 2017; Jurado & Rosselli, 2007; Müller & Kerns, 2015), a summary of latent variable research will further elucidate the developmental sequence through which executive functions arise. Lastly, the identification of evidence-based factor models can inform the hypothesized structure of new test batteries to measure executive functions for implementation into either research or clinical practice.

The current study aimed to (a) determine the empirical support for measurement models of executive functions proposed by past researchers, (b) identify the number of purported executive functions supported by confirmatory factor analyses in the current literature, and (c) determine which published measurement model best fits summary data across studies. To fulfill the first two aims, the current study involved a broad systematic review of research reporting confirmatory factor analyses on batteries of performance-based tasks evaluating executive functions, summarizing both the frequency of model solutions (e.g., unidimensional, three-factor, nested factor models) and the rate at which different factors were included in accepted measurement models (e.g., inhibition, updating, shifting). Considering the significant heterogeneity between the measurement models evaluated by past researchers, the approach to the third aim required a narrower focus on comparable studies, and ultimately considered only those studies assessing the most frequently evaluated factor model within the published literature: the three-factor measurement model of inhibition, shifting, and updating/working memory (Miyake et al., 2000), with updating and working memory merged into updating/working memory because these terms are often used interchangeably in latent variable research. The results of these comparable studies were re-analyzed and fitted to competing factor solutions based on the published literature. The approach of this review was guided by data rather than theory, summarizing past research findings rather than proposing a new model of executive functions. By fulfilling these aims, the current review described the diversity of existing latent variable research on executive functions and further clarified the strength of empirical evidence behind the most common factor solutions proposed by past researchers.

Method

The report of this systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement (Moher, Liberati, Tetzlaff, Altman, and the PRISMA Group, 2009). This review involved only the qualitative and quantitative re-analysis of summary data from published studies, and such a review is exempt from our internal ethics review process. Prior to the literature search, inclusion criteria were established to identify appropriate articles. For inclusion, articles needed to (a) involve a sample or sub-sample of cognitively healthy participants (i.e., without a neurodevelopmental or neurological disorder known to significantly impact cognitive performance) and (b) report a confirmatory factor analysis of a multidimensional measurement model of executive function. Following this criterion, studies that included multiple factors that could be conceptualized as executive functions, but not directly specified by the authors as dimensions of executive function or a synonymous construct (e.g., executive control) were ineligible. As well, measurement models of solely sub-components of executive function were ineligible (e.g., inhibition, Aichert et al., 2012, Friedman & Miyake, 2004; effortful control, Allan & Lonigan, 2011, 2014; problem solving; Cinan, Özen, & Hampshire, 2013;

Scherer & Tiemann, 2014). Eligible models needed to include (c) a minimum of two indicators, deriving from separate tests, per construct evaluated and (d) only performance-based cognitive or neuropsychological outcomes as indicators for the executive function factor(s) (i.e., studies including biometrics, rating scales or symptom inventories as indicators were deemed ineligible for this review), deriving from (e) at least three separate cognitive or neuropsychological tests (i.e., measurement models evaluating the factor structure of multiple outcomes from a single neuropsychological test were ineligible). Lastly, the articles needed to (f) be published in either a peer-reviewed journal or academic book and (g) be written in the English language. For inclusion in the re-analysis, which synthesized a comparable sub-sample of studies testing the most commonly evaluated measurement model in the literature, the articles needed to meet all aforementioned criteria, but also had to have (h) evaluated a measurement model including factors of inhibition, shifting, and updating (or analogous constructs; e.g., mental set-shifting, switching, working memory, etc.) and (i) provide sufficient summary data for re-analysis (i.e., at least a correlation matrix for all test items included in the model).

Literature Search

The systematic literature search covered dates between January 1998 and November 2016. The lower bound of this data range was designated to capture articles following the publication of Miyake et al. (2000) and any articles published just prior to this study that may have involved a confirmatory factor analysis of tests of executive functions. The electronic search strategy involved online searches of the following databases, with search restrictions in parentheses: PsycInfo (Publication type – Peer-reviewed journals, All books; Methodology – Empirical studies, Quantitative studies; Population group – Human; Language – English), PsycArticles (Publication type – Empirical studies, Quantitative studies; Population group – Human), MedLine (Publication type – Journal article; Population group – Human; Language – English), and CINAHL (Publication type – Journal article, Book, Book chapter, Research, Statistics; Language – English). The search protocol involved the following Medical Subject Headings (MeSH), Psychological Index Terms (Tuleya, 2009), and search terms:

> ((MM “Factor Analysis” OR MM “Factor Structure” OR MM “Goodness of Fit” OR MM “Structural Equation Modeling”) OR (MM “Factor Analysis, Statistical” OR MM “Models, Statistical”) OR (“confirmatory factor analysis” OR “CFA” OR “latent variable”)) AND ((DE “Executive Function” OR DE “Cognitive Control” OR DE “Set Shifting” OR DE “Task Switching” OR MM “self regulation”) OR (MM “Executive Function” OR MM “Inhibition (Psychology)” OR MM “Problem Solving”) OR (“executive function*” OR “self-regulat*”))

All retrieved search results were screened twice to ensure that no study went overlooked (Edwards et al., 2002). Following the electronic search, reference lists from peer-reviewed journals were manually searched over the course of data extraction and manuscript preparation, identifying any articles missed by the electronic search protocol (See Figure 1, for a flow diagram of the systematic review process along with the number of articles identified).

Data Extraction

Two independent reviewers extracted relevant information from each article through use of a common data collection spreadsheet. Both reviewers extracted variables related to study characteristics (i.e., authorship, year of publication), sample characteristics (i.e., percent female, mean age, mean years of education, ethnic composition), model characteristics (i.e., name of dependent variables and respective factors), and factor analytic results for accepted measurement models (i.e., χ^2 value and respective *p*-value; comparative fit index, CFI; root mean square error of approximation, RMSEA). For samples eligible for the re-analysis, summary data necessary for a re-analysis of the measurement model was also extracted (i.e., sample size, means/standard deviations, correlation/covariance matrix).

To quantify study quality, reviewers rated articles based on a scale developed specifically for the current review. The majority of confirmatory factor analytic studies involve observational research designs with one time point of data collection (Willoughby et al., 2014), which represents one of the lowest levels of scientific evidence (OCEBM Levels of Evidence Working Group, 2011). Few instruments for rating the quality of this level of research exist in the literature (Sanderson, Tatt, & Higgins, 2007; Vandenbroucke et al., 2007). In turn, the current systematic review strategy applied eleven criteria to rate study quality. These criteria were based largely on standard publication practices for factor analyses (Schreiber, Nora, Stage, Barlow, & King, 2006), with each item scored as either met (1 point) or not met (0 points) and summed for a total study quality score (range: 0–11). The study quality rating scale included the following items:

> (1) the researchers reported a sample size with $\hat{\pi} \geq .80$ to reject the null hypothesis (RMSEA ≥ .05) for a model obtaining a perfect RMSEA (Hancock, 2006), (2) listed at least two demographic variables for each sample evaluated (e.g., mean age, gender composition), (3) indicated that data screening/cleaning for outliers or data transformations to ensure normality was conducted, (4) provided a path diagram of at least one measurement model evaluated or a structural model including all variables from the accepted measurement model, (5) reported the results of a χ^2 goodness-of-fit test and at least two alternative fit indices (e.g., RMSEA, CFI, etc.), (6) listed all of the loadings and (7) residuals for at least one measurement model or structural model evaluated, (8) provided inter-factor correlations (or covariances) for at least one of the multidimensional measurement models or structural models evaluated (if constrained to zero, the authors reported this constraint in the manuscript), (9) reported the means and standard deviations for all manifest variables included in the measurement model, (10) provided a correlation or covariance matrix including all manifest variables included in the measurement model, and (11) had at least three indicators loading on each latent factor in every measurement model evaluated (Roberts & Grover, 2009).

The selection of the power criterion in this scale was based on post-hoc power analyses for model fit that were calculated based of previously published criteria. A power ($\hat{\pi}$) cutoff of ≥ .80 was selected as a conventional threshold in power analysis (Cohen, 1992). Hancock (2006) provides tables to calculate post-hoc power to reject the null hypothesis (i.e., RMSEA ≥ .05) based on three RMSEA values (.00, .02, .04). The tables for the perfect

RMSEA value (i.e., .00) were used to determine whether models met sufficient power (i.e., $\hat{\pi} \geq .80$) because (a) many studies reported perfect RMSEA values and (b) these tables list the smallest required sample sizes to meet this threshold. Stricter thresholds would have resulted in few or no studies meeting this criterion.

Re-Analysis

All articles eligible for the re-analysis provided a correlation matrix for their test battery (included in the Supplementary Materials) and tested the same three-factor model, including factors of inhibition, updating, and shifting or analogous constructs. One study included in the re-analysis (Hedden & Yoon, 2006) reported two factors that could be considered inhibition-related factors (i.e., prepotent response inhibition and resistance to proactive interference). Because prepotent response inhibition was most analogous to the inhibition factor included in other measurement models also eligible for the re-analysis, this factor was included as the inhibition factor in all models run using the correlation matrix for this study, while the resistance to proactive interference factor was left out.

The re-analysis involved two primary aims that rationalized the methodological approach. First, not all researchers examined all factor models supported by the literature with their dataset, and a re-analysis specifying multiple possible measurement models would determine if a specific factor model tended to fit best across published samples. Second, the risk for publication bias was of concern, because most publications identified in the systematic review reported small samples and excellent-fitting models that converged without any errors.

The correlation matrix was re-analyzed by specifying seven alternative measurement models: a unidimensional model, three two-factor models that merged two of the first-order factors (i.e., inhibition = updating; updating = shifting; inhibition = shifting), a three-factor model (i.e., inhibition, updating, and shifting), a nested factor model (i.e., a common executive function bifactor, with shifting-specific and updating-specific factors co-loading on their select indicators and no inhibition-specific factor), and a bifactor model (i.e., a common executive function bifactor with specific factors for inhibition, shifting, and updating). See Figure 2 for a visual representation of each model. Six of these seven models (i.e., all but the bifactor model) were identified as published factor solutions by at least one study in the systematic review. While the full bifactor model was not accepted by any researchers, it was tested as a comparison point for the nested factor model (as done originally by Friedman et al., 2008), permitting evaluation of whether the removal of the inhibition-specific factor improved the fit of the model.

The re-analysis was conducted through a parametric bootstrap simulation based on the published correlation matrix where the data from each study were assumed to be multivariate normal with the observed correlation matrix considered equivalent to the population correlation matrix. For each sample, correlation matrices were computed for 5,000 simulated datasets of equal sample size to that of the original study. For all 5,000 correlation matrices, each factor model was fit to the data. Fit indices were calculated for models that "properly converged," which means the model converged without any errors that would indicate a solution was inadmissible or the estimates were not trustworthy (e.g., a

correlation larger than absolute 1.0, negative residual variances, a non-positive definite latent variable covariance matrix). Throughout the rest of this manuscript, the terms properly converged and converged will be used synonymously. For all samples that properly converged, the CFI and RMSEA were calculated. All factor variances were fixed to 1.0 to set the metric for the factor, and all loadings were freely estimated for all models, with one exception: models with only two indicators on any specific factor in the bifactor or nested factor models had the loadings for those indicators set to be equal for purposes of model identification, as done by previous researchers (Canivez, 2014; Watkins, 2010). The bootstrap re-analysis was conducted in R (R Core Team, 2013), with all factor models fit using the Lavaan package (Rosseel, 2012). The bootstrapping method was validated by testing the accepted models for the Wechsler Adult Intelligence Scale, Fourth Edition (WAIS-IV; Wechsler, 2008) and the Wechsler Intelligence Scale for Children, Fifth Edition (WISC-V; Wechsler, 2014) using published correlation matrices available in the technical manuals for these tests. More details on this validation, along with the results, are provided in the Supplementary Materials.

Model Fit Interpretation—Model fit was evaluated by use of the CFI and RMSEA. These fit indices were selected for three reasons. First, they are commonly reported in the executive function literature, which is why they were included as extracted data elements for the systematic review. The majority of eligible studies reported these fit indices, and researchers within this field are familiar with their use. Second, they are not sensitive to sample size (Fan, Thompson, & Wang, 1999), which was important because the sample sizes varied substantially between studies. And third, they provide a common metric that is comparable across models and offer standard thresholds for acceptable fit and cutoff criteria when comparing alternative models. The RMSEA was also a good choice because it favors parsimony (Hooper, Coughlan, & Mullen, 2008), which was meaningful when comparing models that ranged from simple unidimensional models to those with far more estimated parameters, such as the bifactor model.

Lenient and strict thresholds for acceptable fit and cutoffs for model comparisons in fit were used to guide model acceptance and selection for both the CFI and RMSEA. For the CFI, the lenient and strict thresholds for acceptable fit were ≥ .90 (Bentler & Bonett, 1980) and ≥ .95 (Hu & Bentler, 1999), respectively; and for the RMSEA, the lenient and strict thresholds were ≤.08 and ≤.05, respectively (Browne & Cudeck, 1993). Both the CFI and RMSEA also have cutoffs for significant improvements in model fit when comparing competing models. The lenient and strict cutoffs for change in CFI (i.e., ΔCFI) were ≥.005 and ≥.010, respectively, while the lenient and strict cutoffs for change in RMSEA (i.e., ΔRMSEA) were ≤−.010 and ≤−.015, respectively (Chen, 2007).

The simulated data were interpreted in two ways. The first interpretation evaluated the rate of model acceptance, meaning the percent of bootstrapped models that both converged and met lenient and strict cutoffs for the CFI and RMSEA. Across studies, the means and medians of percent convergence, percent meeting fit thresholds, and percent both converging and meeting fit thresholds (i.e., the rate of model acceptance) were calculated. These percentages were taken to identify the frequency at which a researcher with data from a

battery of executive function tests would (a) have their proposed model converge without any errors that would affect inference and (b) meet standard fit criteria.

The second interpretation evaluated the comparable preference for each model through direct comparisons in fit between competing models. The models were arranged hierarchically based on parsimony for model comparisons, from highest to lowest model complexity: bifactor, nested factor, three-factor, two-factor models (i.e., three different models, all equally parsimonious), and one-factor. For each bootstrapped sample, each model was directly compared to all other models evaluated based on lenient and strict cutoffs for ΔCFI and ΔRMSEA. If a model presented significantly better fit based on a cutoff, it was preferentially selected over an alternative model. If the differences in ΔCFI or ΔRMSEA did not exceed the cutoff, then the more parsimonious model was preferentially selected. If the models were equivalently parsimonious (i.e., the two-factor models), whichever model had the best fit based on absolute CFI or RMSEA was preferentially selected. The results of these analyses were interpreted based on (a) the percent of bootstrapped samples where the model properly converged and was selected based on the ΔCFI or ΔRMSEA cutoffs (hereafter referred to as percent model selection), and (b) the percent of bootstrapped samples where the model was selected based on the ΔCFI or ΔRMSEA cutoffs among only those samples where the model properly converged (hereafter referred to as percent contingent model selection). Across studies, the means and medians of the percent model selection and percent contingent model selection were taken.

The percent model selection summarizes the frequency at which a researcher with data from a battery of executive function tests would have a model converge and select that model over competing models. The percent contingent model selection summarizes the frequency at which a researcher would select a model among only those samples where that model properly converged (i.e., in samples where that model converges, how often it has superior fit to competing models). The comparison between models was made regardless of whether or not the models met standard fit thresholds. In turn, even if a model is selected over other models with a high frequency, the model does not necessarily meet the conventional fit thresholds used to interpret rates of model acceptance (i.e., CFI ≥ .90/.95; RMSEA ≤ .05/.08). In turn, the percent model acceptance and model selection must be interpreted in combination.

Results

Systematic Review

The literature review identified 40 articles meeting eligibility criteria for the systematic review reporting measurement models for 46 different samples (see Figure 1). Among those eligible studies, 17 articles provided sufficient data for the re-analysis of 21 samples. A reference list of full-text articles reviewed during the literature search, but ultimately not included in the systematic review, is provided in the Supplementary Materials along with a reason for their exclusion.

A large set of studies examined for the current review pulled participants from the Victoria Longitudinal Study (de Frias, Dixon, & Strauss, 2006, 2009; McFall et al., 2013, 2014;

Sapkota, Vergote, Westaway, Jhamandas, & Dixon, 2015; Thibeau, McFall, Wiebe, Anstey, & Dixon, 2016), the Colorado Longitudinal Twin Study (Friedman et al., 2006, 2007, 2008, 2009, 2011, 2016), and the Family Life Project study (Willoughby, Blair, & The Family Life Project Investigators, 2016; Willoughby, Blair, Wirth, Greenberg, & The Family Life Project Investigators, 2010; Willoughby et al., 2012a; Willoughby, Wirth, Blair, & The Family Life Project Investigators, 2012b) with definitive or potential overlap among the participants included in their analyses. Some cross-sectional studies also reported analyses for the same participant data across different articles (Miller, Giesbrecht, Müller, McInerney, & Kerns, 2012; Miller, Müller, Giesbrecht, Carpendale, & Kerns, 2013; van der Ven et al., 2012, 2013; Usai, Viterbori, Traverso, & De Franchis, 2014; Viterbori, Usai, Traverso, & De Franchis, 2015; Rose, Feldman, & Jankowski, 2011, 2012). To avoid representing the same participants twice in the review, the studies involving the largest samples and the most executive function tasks were ultimately included in the systematic review and re-analysis (de Frias et al., 2009; Friedman et al., 2011; Miller et al., 2012; Rose et al., 2012; van der Ven et al., 2013, Willoughby et al., 2012a).

Most studies reporting confirmatory factor analyses on executive functions involved cross-sectional research designs; and for the limited amount of longitudinal studies identified, only one wave of measurement per study was represented in the current review and re-analysis. For one longitudinal study evaluating the same battery of executive function tasks at multiple time points, the data from the first wave were considered for the current review and re-analysis (i.e., de Frias et al., 2009). The consideration of just the first wave data made the study design more comparable to other studies in the review; however, in contexts where the task battery changed, the wave with the most available executive function tasks or the most complete summary data was considered in the current review (i.e., Willoughby et al., 2012a; Lee et al., 2013).

Qualitative Synthesis

Demographics of samples evaluated—Table 1 provides the demographic characteristics for each sample included in the systematic review along with an estimate of study quality. Among the samples reported by studies included in the systematic review, 9 samples (n = 2,614; $\bar{x}$ % female = 49.81%) consisted of preschool aged children ($\bar{x}$ age range: 3.01 to 5.77 years), 15 samples (n = 2,374; $\bar{x}$ % female = 48.54%) consisted of school-aged children ($\bar{x}$ age range: 6.42 to 11.88 years), 3 samples (n = 1,040; $\bar{x}$ % female = 48.87%) consisted of adolescents ($\bar{x}$ age range: 14.41 to 17.30 years), 9 samples (n = 2,070; $\bar{x}$ % female = 51.27%) consisted of adults ($\bar{x}$ age range: 19.75 to 25.70 years), and 8 samples (n = 1,112; $\bar{x}$ % female = 61.44%) consisted of older adults ($\bar{x}$ age range: 60.24 to 74.40 years). Two studies evaluated samples with participants spanning multiple age groups (n = 546), including a child to young adult sample ($\bar{x}$ age range: 7.20 to 20.80 years; Huizinga et al., 2006) and a merged young and older adult sample ($\bar{x}$ age range: 21.00 to 71.00 years; Pettigrew & Martin, 2014). Overall, 9,756 participants ($\bar{x}$ % female = 52.56%) were represented in the systematic review.

Among the 18 samples with some race or ethnicity information provided, 10 samples were predominantly White, 3 samples were majority non-White, and 5 samples were identified as

ethnically Chinese (Lee et al., 2012; Xu et al., 2013) or from Chinese schools (Duan et al., 2010). Study quality was on average 8.32 (SD = 1.91; range: 1 to 11) across age groups. It was similar on average for preschool children ($\bar{x}$ = 8.56), school-aged children ($\bar{x}$ = 8.31), adolescents ($\bar{x}$ = 8.00), and adults ($\bar{x}$ = 9.22). It was lower for older adults ($\bar{x}$ = 6.86) due to one study receiving a single study quality point (Frazier et al., 2015). When this outlier was removed, the mean study quality for older adult studies increased to 7.83, which was more similar to the other age bands.

Model fit indices and accepted models—Table 2 provides fit indices for accepted measurement models identified by the systematic review, along with estimated power (based on N and df; Hancock, 2006), the number of factors, and names of factors included in the accepted model. Considering fit indices, all accepted models had CFI values ≥ .95 and all RMSEA values ≤ .06 (with the exception of one study with CFI = .92; McVay & Kane, 2012), indicating excellent statistical fit for the models (Hu & Bentler, 1999). These excellent model fit statistics stood in contrast to the predominantly low power estimates across studies, which came to an average of 0.44 (SD = 0.32; range = 0.08 to 0.99). The accepted models included anywhere between one to five factors. Overall, 8 studies accepted a one-factor model (17.39%), 18 accepted a two-factor model (39.13%), 14 accepted a three-factor model (30.43%), 1 accepted a four-factor model (2.17%), 1 accepted a five-factor model (2.17%), and 4 accepted a nested factor model (8.70%). For the calculation of these totals and those reported below, Carlson et al. (2014) was considered to have accepted a one-factor model based on parsimony, although these authors specified no preference between a one-factor or two-factor model; and de Frias et al. (2009) accepted a two-factor model for their Cognitively Normal Subsample, although this model was never formally evaluated.

For preschool samples, roughly half of researchers accepted a one-factor model solution (Number of studies [k] = 5; 55.56%; Carlson et al., 2014; Masten et al., 2012; Wiebe et al., 2008; Wiebe et al., 2012; Willoughby et al., 2012a), while the other half preferred a two-factor solution (k = 4; 44.44%; Lerner & Lonigan, 2014; Miller et al., 2012; Monette et al., 2015; Usai et al., 2014). Among the school-aged samples, the most commonly accepted model was the three-factor model (k = 7; 46.67%; Agostino et al., 2010; Arán-Filippetti, 2013, Duan et al., 2010; Lambek & Shevlin, 2011; Lehto et al., 2003; Rose et al., 2012), while a smaller set of studies supported a two-factor (k = 4; 26.67%; Brocki & Tillman, 2014; Lee et al., 2012, 2013; van der Ven et al., 2013) or one-factor solution (k = 3; 20%; Brydges et al., 2012; Xu et al., 2013). One study involving a school-aged sample supported a model best categorized as a nested factor model (k = 1; 6.67%; van der Sluis et al., 2007), although these researchers did not label it as such. Among the three adolescent studies, researchers reported a single nested factor model (k = 1; 33.33%; Friedman et al., 2011) and a pair of three-factor models (k = 2; 66.66%; Lambek & Shevlin, 2011; Xu et al., 2013). For the adult studies, the support was roughly split between a two-factor model (k = 3; 33.33%; Klauer et al., 2010; McVay & Kane, 2012; Was, 2007), a three-factor model (k = 2; 22.22%; Klauer et al., 2010; Miyake et al., 2000), and a nested factor model (k = 2; 22.22%; Fleming et al., 2016; Ito et al., 2015). One study supported a four-factor model (k = 1; 11.11%; Chuderski et al., 2012) and another supported a five-factor model (k = 1, 11.11%; Fournier-Vicente et al., 2008). The older adult samples predominantly supported a two-factor model

(*k* = 5, 62.5%; Bettcher et al., 2016; de Frias et al., 2009; Frazier et al., 2015; Hedden & Yoon, 2006; Hull et al., 2008), while a smaller, but substantial percentage supported a three-factor model (*k* = 3, 37.5%; Adrover-Roig et al., 2012; de Frias et al., 2009; Vaughan & Giovanello, 2010).

Table 3 provides counts and frequencies of how often a specific construct was represented in an accepted factor model. The most common factors were those included in the original measurement model by Miyake and colleagues (2000), with updating/working memory (*k* = 33; 71.74% of models) being the most frequent, followed by inhibition (*k* = 24; 52.17%), and then by shifting (*k* = 20; 43.48%). A small number of studies merged these factors, including inhibition and shifting (*k* = 5; 10.87%), inhibition and updating/working memory (*k* = 1; 2.17%), and shifting and updating/working memory (*k* = 3; 6.52%). Two studies included factors of strategic retrieval or access to long-term memory (*k* = 2; 4.35%; Adrover-Roig et al., 2012; Fournier-Vicente et al., 2008).

Some differences occurred in terms of the factors represented across age spans. A global executive function factor was represented among 23.91% of models (*k* = 11), but constituted a unidimensional factor among children and a nested bifactor among adolescents and adults. No sample beyond the school-aged years provided a unidimensional model solution, and a global executive function factor was not observed among any eligible older adult samples. No preschool sample identified shifting as a separate factor, while all three factors were represented in all groups above 6 years of age.

Tests used as indicators—In the Supplementary Materials, Tables S1 and S2 list the indicators organized by factors for child/adolescent and adult studies, respectively. The division between child/adolescent and adult samples was set at a mean age of 16 years, where those with a mean age at or below 16 years were considered child/adolescent (*k* = 21) and those with a mean age over 16 years were considered adult (*k* = 18). Few studies involved the same battery of tests for all indicators evaluated, but a small number of measures were common in the evaluation of specific constructs. The tests below are categorized based on either task or paradigm, and do not necessarily indicate that the studies were using the exact same task or the exact same dependent variable derived from that task. In some contexts, the exact same task or a highly similar task was used across studies (e.g., Digit Span Backward); however, in other contexts, a similar paradigm was used to guide the design of similar, but distinguishable tasks. For example, the Stroop paradigm among children comes in multiple different varieties of tasks, including a Boy-Girl Stroop, Day-Night Stroop, and Color-Word Stroop; all of which involve different stimuli, but similar task demands, and they load onto inhibition.

The most frequent indicator of inhibition for child/adolescent studies were tasks using the Stroop paradigm (*k* = 11), followed by tasks using the Go/No-go paradigm (*k* = 7). Tasks using a Tower paradigm were the third most common indicator for inhibition among child/adolescent studies (*k* = 4). The most commonly used indicator for updating/working memory was the Digit Span Backward task (*k* = 7), followed by the Letter-Number Sequencing task (*k* = 3) and tasks using the *n*-back paradigm (*k* = 3). For shifting, tasks with card sorting paradigms were the most commonly used as indicators (*k* = 6), while tasks

using a Trail Making paradigm were the second most commonly used ($k = 5$) and tasks using a verbal fluency paradigm were the third most commonly used ($k = 4$).

In terms of adult studies, there was a greater frequency at which specific measures were used as indicators across studies. For inhibition, a substantial portion of the adult studies used tasks involving a Stroop paradigm ($k = 16$), followed by an Antisaccade task ($k = 11$), and then a Stop-Signal task ($k = 7$). For updating/working memory, the most frequently used indicators were tasks using the *n*-back paradigm ($k = 8$) and the Letter Memory task ($k = 8$), followed by the Keep Track task ($k = 6$) and Digit Span Backwards task ($k = 5$). The measurement of shifting was more variable, but still a substantial portion of researchers used the Number-Letter task ($k = 10$), followed by the Plus-Minus task ($k = 5$) and the Local Global task ($k = 4$).

The data extraction protocol involved the extraction of the task names, and did not focus on the specific dependent variables derived from each of these tasks that were ultimately included in measurement models. A post-hoc evaluation explored the variety of scores that different researchers used in their models for the most commonly used paradigm: the Stroop task as an indicator for inhibition. The Stroop task consists of congruent/neutral conditions along with incongruent conditions. In congruent/neutral conditions, participants read color words (e.g., blue, red) written in either black ink or their corresponding ink color, or they named the ink color of a non-verbal stimulus (e.g., a line of asterisks or X's). In the incongruent condition, participants see color words written in incongruent ink colors (e.g., blue written in red ink) and they are asked to read the ink color, inhibiting the automatic response of reading the word. Among children, similar tasks use alternative stimuli, such as the Day-Night Stroop where children are shown a sun or moon and asked to say night or day, respectively.

Among the 11 child/adolescent studies using a Stroop-like task, 7 studies included a Stroop Color-Word paradigm, while the remainder involved Day-Night, Boy-Girl, or other Stroop-like task. Within the 7 studies using the color-word approach, 6 different dependent variables were identified, including the difference in time-to-completion between the incongruent and neutral/congruent conditions (Agostino et al., 2010; Brydges et al., 2012), the total number correct in the incongruent condition (Arán-Filippetti, 2013), the difference in the number of correct responses between the incongruent and neutral/congruent conditions (Brocki et al., 2014), the median response latency on incongruent trials (Huizinga et al., 2006), the number of items named per second (van der Sluis et al., 2007), and the reaction time difference between incongruent and neutral/congruent conditions (Xu et al., 2013).

Among the 16 studies using a Stroop paradigm among adult samples, 6 different dependent measures were derived from the same test, including a reaction time difference score between incongruent and neutral/congruent conditions (Fleming et al., 2016; Friedman et al., 2011; Fournier-Vicente et al., 2008; Hull et al., 2008; Ito et al., 2015; Klauer et al., 2010; Miyake et al., 2000; Was, 2007), a ratio of proportion correct in the incongruent condition to proportion correct in the neutral/congruent condition (Chuderski et al., 2012), an interference index (de Frias et al., 2009), the total correct in the incongruent condition statistically controlling for the total correct in the neutral/congruent condition (Bettcher et

al., 2016; Frazier et al., 2015; Pettigrew et al., 2014), the reaction time for correct incongruent trials (Vaughan & Giovanello, 2010), and the reaction time for incongruent trials regardless of accuracy (McVay & Kane, 2012).

Bootstrapped Re-Analysis

As noted earlier, a total of 21 samples met eligibility criteria for the re-analysis. These samples were not evenly divided between the age bands used to categorize the studies in the qualitative synthesis: preschool ($k = 2$), school-age ($k = 8$), adolescent ($k = 2$), adult ($k = 5$), and older adult ($k = 4$). Due to the wide span of ages, the samples were stratified into two samples with 16 years of age as the cut point, where 10 samples were considered adult (i.e., >16 years of age) and 11 samples were considered child and adolescent (i.e., ≤16 years of age). Among the child/adolescent studies, the choice was made to exclude the 2 re-analyzed preschool samples from the calculation of summary statistics for that age range (e.g., mean/median percent convergence, mean/median percent meeting fit criteria). This decision was based on (a) the observation that no separate shifting factor was observed for preschool samples in the qualitative synthesis, (b) the extensive literature detailing the early childhood years as unique and fundamental for executive function development (Müller & Kerns, 2015), and (c) the conceptualization of shifting as an ability that arises later in executive function development (Garon et al., 2008). The exclusion of the preschool samples led to 9 child/adolescent samples with an average age span ranging from 8.33 to 14.41 years. The age span for the adult studies ranged from 17.30 to 72.24. The 17-year-old sample (Friedman et al., 2011) was included with the other adult sample due to factor analytic research observing stability of the structure of executive functions from this age into early adulthood (Friedman et al., 2016). Older adults were included within this age band because (a) there was an insufficient number older adult studies to compose its own group; and (b) although there is evidence for age-related declines in performances on executive function tasks (Reynolds & Horton, 2008), the qualitative findings did not provide definitive evidence for de-differentiation. Unlike the preschool age band, all three constructs were represented among this age group, and the oldest sample evaluated produced a three-factor solution (Vaughan & Giovanello, 2010).

Percent convergence—Provided in the Supplementary Materials, Tables S3 and S4 list the percentage of models that converged among the 5,000 bootstrapped samples for each measurement model specified for child/adolescent and adult studies, respectively. The percent convergence is presented for each individual study, and a mean and median percent convergence is presented for all studies. These summary statistics for percent convergence are visually presented in Figures 3a and 4a for child/adolescent and adult studies, respectively. For both the child/adolescent and adult studies, the rates of convergence were related to model complexity, where models with more parameters tended to properly converge less often; however, the more complex set of models differed across age spans in terms of their frequency of convergence. For example, among adult studies, there was a clear negative relationship between percent convergence and model complexity. The bifactor model converged the least often ($\bar{x} = 24\%$; $Mdn = 10\%$). The nested factor ($\bar{x} = 57\%$; $Mdn = 53\%$) and three-factor models ($\bar{x} = 45\%$; $Mdn = 40\%$) converged infrequently and less often than the three two-factor models, which all converged at roughly the same rate: inhibition-

shifting merged ($\bar{x}$ = 76%; *Mdn* = 86%), inhibition-updating merged ($\bar{x}$ = 71%; *Mdn* = 77%), and shifting-updating merged ($\bar{x}$ = 65%; *Mdn* = 66%). The unidimensional model converged for almost every bootstrapped sample ($\bar{x}$ = 95%; *Mdn* = 99%).

In contrast to the adult studies, the frequency of convergence among the child/adolescent samples was slightly different, where the model that converged the least often was the three-factor model ($\bar{x}$ = 36%; *Mdn* = 26%), while the nested factor ($\bar{x}$ = 59%; *Mdn* = 60%) and bifactor models ($\bar{x}$ = 48%; *Mdn* = 49%) converged at closer frequencies. For the three two-factor models, the models merging the shifting factor tended to converge more often. The inhibition-shifting merged ($\bar{x}$ = 76%; *Mdn* = 89%) and shifting-updating merged models ($\bar{x}$ = 71%; *Mdn* = 56%) converged more often than the inhibition-updating merged model ($\bar{x}$ = 59%; *Mdn* = 55%). As with the adult studies, the unidimensional model converged for almost every bootstrapped sample ($\bar{x}$ = 97%; *Mdn* = 100%).

Percent of converged models meeting fit criteria—Tables S3 and S4, in the Supplementary Materials, list the percentage of the converged models that met lenient and strict fit thresholds for each measurement model specified for child/adolescent and adult studies, respectively. The trend in terms of meeting fit thresholds was generally in the opposite direction of model convergence, where the more complex models tended to fit better than the simpler models. This was true for both the CFI and RMSEA, and the trend is visually represented in Figures 3b and 4b for child/adolescent and adult studies, respectively. As also clearly demonstrated by these figures, the strict fit thresholds were rarely met for most models, whereas the lenient fit thresholds, though met more often, were still met infrequently.

For the adult studies, the bifactor model met lenient (CFI: $\bar{x}$ = 63%; *Mdn* = 55%; RMSEA: $\bar{x}$ = 61%; *Mdn* = 60%) and strict fit criteria (CFI: $\bar{x}$ = 36%; *Mdn* = 30%; RMSEA: $\bar{x}$ = 25%; *Mdn* = 25%) the most often among the bootstrapped samples for which this model converged. The nested factor model met lenient (CFI: $\bar{x}$ = 54%; *Mdn* = 52%; RMSEA: $\bar{x}$ = 59%; *Mdn* = 58%) and strict fit criteria (CFI: $\bar{x}$ = 23%; *Mdn* = 18%; RMSEA: $\bar{x}$ = 18%; *Mdn* = 14%) at roughly the same rate that the three-factor model met lenient (CFI: $\bar{x}$ = 48%; *Mdn* = 44%; RMSEA: $\bar{x}$ = 57%; *Mdn* = 57%) and strict fit criteria (CFI: $\bar{x}$ = 19%; *Mdn* = 10%; RMSEA: $\bar{x}$ = 16%; *Mdn* = 15%). The two-factor models all met the fit criteria at about the same frequency, although the inhibition-updating merged model met the ≤.08 RMSEA criterion ($\bar{x}$ = 42%; *Mdn* = 45%) at a greater rate than the other two-factor models, as made visually evident by a peak in the forest plot line in Figure 4b.

For the child/adolescent studies, the bifactor met lenient (CFI: $\bar{x}$ = 64%; *Mdn* = 71%; RMSEA: $\bar{x}$ = 50%; *Mdn* = 52%) and strict fit criteria (CFI: $\bar{x}$ = 39%; *Mdn* = 42%; RMSEA: $\bar{x}$ = 21%; *Mdn* = 21%) the most often among the bootstrapped samples for which this model converged. The three-factor model tended to meet lenient and strict fit criteria at about the same frequency as the nested factor model. Similarly, the two-factor models all tended to meet lenient and strict fit criteria at roughly the same rate, while the unidimensional model met lenient (CFI: $\bar{x}$ = 36%; *Mdn* = 48%; RMSEA: $\bar{x}$ = 32%; *Mdn* = 21%) and strict fit criteria (CFI: $\bar{x}$ = 11%; *Mdn* = 6%; RMSEA: $\bar{x}$ = 11%; *Mdn* = 5%) the least often.

The percent of converged samples meeting fit thresholds cannot be properly understood without appreciating the percent of models converging among the bootstrapped samples. Those models that did not converge did not provide fit indices to contribute to this overall estimate, indicating that the percent of fitting models based on fit thresholds alone may overestimate how often these models were accepted among the 5,000 bootstrapped samples. In turn, the next section presents how often models both converged and met fit criteria among the 5,000 bootstrapped samples across studies.

Rate of model acceptance based on percent of models both converging and meeting fit criteria—Among the 5,000 bootstrapped samples for each study, the frequency at which models both converged and met fit criteria was quite low across different models estimated, although some models tended to be accepted more often than others. The percent of samples for which a specified model both converged and met fit criteria is provided for multiple fit thresholds in Tables S3 and S4 within the Supplementary Materials for child/adolescent and adult samples, respectively. Figures 3c and 4c offer a visual representation of these values. These values constitute the percent of samples in which this model would be accepted by a researcher, in that the model both properly converged and met criteria indicative of good model fit.

Among the adult studies, the rate at which models were deemed acceptable was quite low based on lenient fit criteria and extremely low based on strict fit criteria. The nested factor model was the most often accepted model based on both the lenient (CFI: $\bar{x}$ = 41%; *Mdn* = 26%; RMSEA: $\bar{x}$ = 42%; *Mdn* = 27%) and strict fit indices (CFI: $\bar{x}$ = 17%; *Mdn* = 10%; RMSEA: $\bar{x}$ = 13%; *Mdn* = 6%). Based on lenient fit indices, the three-factor model was the second most often accepted model (CFI: $\bar{x}$ = 25%; *Mdn* = 13%; RMSEA: $\bar{x}$ = 32%; *Mdn* = 19%); however, based on strict fit indices, the bifactor model (CFI: $\bar{x}$ = 11%; *Mdn* = 4%; RMSEA: $\bar{x}$ = 8%; *Mdn* = 3%) was accepted at about the same frequency as the three-factor model (CFI: $\bar{x}$ = 8%; *Mdn* = 5%; RMSEA: $\bar{x}$ = 7%; *Mdn* = 4%). The two-factor models did not differ from the three-factor model or each other in how often they were accepted based on strict fit criteria; however, based on lenient fit criteria, the inhibition-updating merged model was the most often accepted of the two-factor models (CFI: $\bar{x}$ = 19%; *Mdn* = 10%; RMSEA: $\bar{x}$ =36%; *Mdn* = 31%). The acceptance rate based on RMSEA was slightly higher for this model compared to the three-factor model, but the three-factor model was accepted more often based on CFI. The unidimensional model was comparable to the two-factor models in terms of strict fit criteria, and was very rarely accepted based on lenient fit criteria as well (CFI: $\bar{x}$ = 8%; *Mdn* = 0%; RMSEA: $\bar{x}$ = 13%; *Mdn* = 3%).

The child/adolescent studies did not follow the same trend as the adult studies. As clearly presented in Figure 3c, no model stood out as the most often accepted. Instead the inverse occurred, where two models were more frequently *not* accepted, specifically – based on lenient fit criteria – the inhibition-updating merged model (CFI: $\bar{x}$ = 20%; *Mdn* = 20%; RMSEA: $\bar{x}$ = 13%; *Mdn* = 12%) and the three-factor model (CFI: $\bar{x}$ = 21%; *Mdn* = 10%; RMSEA: $\bar{x}$ = 11%; *Mdn* = 8%) rarely converged and met fit thresholds. Based on lenient fit criteria, there was no clear delineation between the unidimensional (CFI: $\bar{x}$ = 36%; *Mdn* = 48%; RMSEA: $\bar{x}$ = 32%; *Mdn* = 21%), shifting-updating merged (CFI: $\bar{x}$ = 35%; *Mdn* = 31%; RMSEA: $\bar{x}$ = 25%; *Mdn* = 32%), inhibition-shifting merged (CFI: $\bar{x}$ = 34%; *Mdn* =

32%; RMSEA: $\bar{x}$ = 27%; *Mdn* = 30%), nested factor (CFI: $\bar{x}$ = 31%; *Mdn* = 26%; RMSEA: $\bar{x}$ = 21%; *Mdn* = 18%), or bifactor models (CFI: $\bar{x}$ = 28%; *Mdn* = 23%; RMSEA: $\bar{x}$ = 20%; *Mdn* = 22%). There was a bit more of a distinction based on strict fit criteria, where the nested factor (CFI: $\bar{x}$ = 17%; *Mdn* = 13%; RMSEA: $\bar{x}$ = 7%; *Mdn* = 4%) and bifactor models (CFI: $\bar{x}$ = 16%; *Mdn* = 13%; RMSEA: $\bar{x}$ = 8%; *Mdn* = 9%) were more often accepted based on CFI, but this trend was not evident based on the RMSEA, which takes model complexity into account.

Model selection based on ΔCFI and ΔRMSEA comparisons—For child/adolescent and adult samples, respectively, Tables S5 and S6 in the Supplementary Materials provide the percent model selection (i.e., the frequency at which a model converged and was selected among 5,000 bootstrapped samples) and the percent contingent model selection (i.e., the frequency at which a model was selected among samples where the model converged). These findings are presented visually in Figures 5 and 6 for child/adolescent and adult samples, respectively.

Among the adult studies, the rate at which models both converged and were selected was quite low. Figure 6a illustrates two peaks around the unidimensional and nested factor models. Based on both the lenient and strict ΔRMSEA cutoffs, which penalizes for model complexity, the unidimensional model showed the highest frequency of model selection (Lenient ΔRMSEA: $\bar{x}$ = 26%; *Mdn* = 16%; Strict ΔRMSEA: $\bar{x}$ = 32%; *Mdn* = 27%). However, based on the ΔCFI cutoffs, the rates of selection of the unidimensional model were much lower (Lenient ΔCFI: $\bar{x}$ = 13%; *Mdn* = 4%; Strict ΔCFI: $\bar{x}$ = 15%; *Mdn* = 5%). The nested factor model was most preferred based on ΔCFI cutoffs (Lenient ΔCFI: $\bar{x}$ = 30%; *Mdn* = 21%; Strict ΔCFI: $\bar{x}$ = 26%; *Mdn* = 20%); however, based on ΔRMSEA cutoffs, the nested factor model was less preferred than many more parsimonious models, including the three-factor model, a pair of two-factor models (i.e., inhibition-updating merged and inhibition-shifting merged), and the unidimensional model.

As shown by a peak in Figure 6b, based on ΔCFI cutoffs, the nested factor (Lenient ΔCFI: $\bar{x}$ = 57%; *Mdn* = 61%; Strict ΔCFI: $\bar{x}$ = 53%; *Mdn* = 59%) and bifactor models (Lenient ΔCFI: $\bar{x}$ = 55%; *Mdn* = 62%; Strict ΔCFI: $\bar{x}$ = 49%; *Mdn* = 52%) were the most frequently selected among samples where those models converged. The ΔRMSEA cutoffs, which penalize for model complexity, did not show this same preference for the nested factor or bifactor models. Based on ΔRMSEA cutoffs, the unidimensional, inhibition-updating merged, inhibition-shifting merged, three-factor, and nested factor models all showed similar frequencies of contingent model selection.

Among the child/adolescent studies, there was a clear peak in Figure 5a based on ΔRMSEA cutoffs, evidencing support for the unidimensional model (Lenient ΔRMSEA: $\bar{x}$ = 46%; *Mdn* = 31%; Strict ΔRMSEA: $\bar{x}$ = 53%; *Mdn* = 43%). For the ΔCFI cutoffs, the peak was not as prominent (Lenient ΔCFI: $\bar{x}$ = 21%; *Mdn* = 9%; Strict ΔCFI: $\bar{x}$ = 26%; *Mdn* = 10%), but still evidenced a higher rate of model selection compared to all other models. In terms of contingent model selection, the results were slightly different. As shown in Figure 5b, there was again a peak based on ΔRMSEA cutoffs, evidencing support for the unidimensional model (Lenient ΔRMSEA: $\bar{x}$ = 46%; *Mdn* = 35%; Strict ΔRMSEA: $\bar{x}$ = 54%; *Mdn* = 43%).

However, the ΔCFI cutoffs, which do not penalize for model complexity, showed a peak in contingent model selection for the nested factor model (Lenient ΔCFI: $\bar{x}$ = 41%; *Mdn* = 38%; Strict ΔCFI: $\bar{x}$ = 42%; *Mdn* = 34%).

Mean fit indices, inter-factor correlations, and inter-item correlations— Available in the Supplementary Materials, Tables S7 and S8 provide the mean fit indices (i.e., CFI and RMSEA) and 95% confidence intervals for child/adolescent and adult studies, respectively. These statistics are based only on the models that converged and provided an estimate of the fit indices. For all models that converged involving correlated factors, Tables S9 and S10 (see Supplementary Materials) for child/adolescent and adult studies, respectively, provide the mean inter-factor correlations and 95% confidence intervals. For studies included in the bootstrap re-analysis, the mean correlations between indicators was also calculated per each construct from the observed correlation matrices. These values are also provided in the Supplementary Materials in Table S11. For updating indicators, the correlations were similar between child/adolescent studies ($\bar{x}$ = 0.41; *Mdn* = 0.30) and adult studies ($\bar{x}$ = 0.38; *Mdn* = 0.33). For shifting indicators, the correlations were also similar between child/adolescent ($\bar{x}$ = 0.29; *Mdn* = 0.33) and adult studies ($\bar{x}$ = 0.30; *Mdn* = 0.26). However, for inhibition indicators, the inter-item correlations were higher for child/adolescent studies ($\bar{x}$ = 0.29; *Mdn* = 0.26) than adult studies ($\bar{x}$ = 0.16; *Mdn* = 0.18).

Post-hoc evaluation of publication bias—The re-analysis focused on rates of model acceptance and selection regardless of which model was originally supported by each individual study. A post-hoc analysis evaluated the presence of publication bias by examining the rate of model acceptance among the 5,000 bootstrapped samples for the model originally accepted by the researchers using their observed sample. This analysis was done using only those studies with accepted models that corresponded to those seven evaluated in the re-analysis, which resulted in 10 child/adolescent samples and 8 adult samples. Although these values are present in Tables S3 and S4, they are presented in isolation in Table S12 as well (see Supplementary Materials) for the convenience of the reader. Among child/adolescent studies, the rate at which the originally accepted models would be accepted among the 5,000 bootstrapped samples was low using both lenient fit criteria (CFI: $\bar{x}$ = 36%; *Mdn* = 43%; RMSEA: $\bar{x}$ = 33%; *Mdn* = 31%) and strict fit criteria (CFI: $\bar{x}$ = 15%; *Mdn* = 15%; RMSEA: $\bar{x}$ = 13%; *Mdn* = 12%). Among adult studies, this rate was also low using lenient (CFI: $\bar{x}$ = 37%; *Mdn* = 14%; RMSEA: $\bar{x}$ = 44%; *Mdn* = 32%) and strict fit criteria (CFI: $\bar{x}$ = 10%; *Mdn* = 5%; RMSEA: $\bar{x}$ = 8%; *Mdn* = 5%).

A similar post-hoc analysis evaluated the frequency of model selection and contingent model selection of the originally supported models reported in published studies, with results summarized in the Supplementary Materials in Table S13. Among child/adolescent studies, the rate at which the originally selected models were preferentially selected among the 5,000 bootstrapped samples was low using both lenient cutoffs (ΔCFI: $\bar{x}$ = 31%; *Mdn* = 33%; ΔRMSEA: $\bar{x}$ = 37%; *Mdn* = 27%) and strict cutoffs (ΔCFI: $\bar{x}$ = 33%; *Mdn* = 33%; ΔRMSEA: $\bar{x}$ = 37%; *Mdn* = 21%). Among adult studies, this rate was also low using lenient (ΔCFI: $\bar{x}$ = 38%; *Mdn* = 34%; ΔRMSEA: $\bar{x}$ = 14%; *Mdn* = 14%) and strict fit cutoffs (ΔCFI: $\bar{x}$ = 34%; *Mdn* = 30%; ΔRMSEA: $\bar{x}$ = 8%; *Mdn* = 7%). In terms of contingent model

selection, the models were selected at a slightly higher rate among those bootstrapped samples where the originally selected model converged, based on lenient (ΔCFI: $\bar{x}$ = 52%; *Mdn* = 56%; ΔRMSEA: $\bar{x}$ = 22%; *Mdn* = 15%) and strict cutoffs (ΔCFI: $\bar{x}$ = 47%; *Mdn* = 53%; ΔRMSEA: $\bar{x}$ = 14%; *Mdn* = 9%).

Discussion

The systematic review and re-analysis summarized an extensive body of research exploring executive functions over the last two decades, identifying a large set of studies producing fairly consistent findings about the structure of executive functions over the course of the lifespan. A qualitative synthesis of this research covered sample demographics, test selection, study quality, model fit, and the frequency at which different constructs and models appeared in the published literature. The existing literature has the appearance of being quite consistent, but that appearance is partially due to overlapping samples across studies and potential publication bias. Complementing the qualitative synthesis, a re-analysis of correlation matrices from a sub-sample of eligible studies compared seven competing measurement models reported in the published literature (see Figure 2), attempting to quantitatively identify a best-fitting measurement model for child/adolescent and adult samples.

Findings from the Qualitative Synthesis

The executive function constructs identified most often included inhibition, updating/working memory, and shifting; however, the number of constructs represented in accepted measurement models varied by the age of the sample evaluated. The majority of samples identified were composed of children and adolescents (k = 27), while a smaller portion of studies involved adults (k = 9) and older adults (k = 8). In terms of the factor models supported by eligible studies, there was evidence for increasing multidimensionality of executive functions over the course of development. Preschool samples were roughly split between a one-factor and two-factor solution, with no studies identifying a specific shifting factor. School-aged samples showed more support for a three-factor model than a two-factor model, while the adolescent samples supported three-factor and nested factor solutions. There was comparable support for two-factor, three-factor, and nested factor models among adult samples. Two of the studies producing a two-factor solution among adults did not test a three-factor solution (McVay & Kane, 2012; Was, 2007), and the other involved two studies and found a three-factor solution in their second study (Klauer et al., 2010). Combined, these findings indicate a gradual differentiation of executive functions from preschool into adulthood, and the potential emergence of a specific shifting factor around school-age to adolescence. This is consistent with some leading theories relating to the neurodevelopment of executive functions (Bardikoff & Sabbagh, 2017; Garon et al., 2008; Müller & Kerns, 2015).

Although consistent with developmental theories, the increased multidimensionality in factor solutions with age could alternatively derive from methodological differences between child and adult studies; specifically, differences in the number of indicators used per construct in measurement models. A close re-examination of Tables S1 and S2 in the

Supplementary Materials indicates a greater frequency of factors with just two indicators for child/adolescent studies in comparison to adult studies. Specifically, among child/adolescent studies, six studies used just two indicators for inhibition (Duan et al., 2010; Lambek & Shevlin, 2011; Lehto et al., 2003; Rose et al., 2012; Usai et al., 2014; Xu et al., 2013), six studies used just two indicators for updating/working memory (Agostino et al., 2010; Duan et al., 2010; Lambek & Shevlin, 2011; Usai et al., 2014; Willoughby et al. 2012a; Xu et al., 2013), and seven studies used just two indicators for shifting (Agostino et al., 2010; Duan et al., 2010; Lehto et al., 2003; Monette et al., 2015; Rose et al., 2012; Usai et al., 2014; Xu et al., 2013). In contrast, among adult studies, three studies used just two indicators for inhibition (de Frias et al., 2009; Frazier et al., 2015; Klauer et al., 2010), two studies used just two indicators for updating/working memory (de Frias et al., 2009; Klauer et al., 2010), and two studies used just two indicators for shifting (de Frias et al., 2009; Frazier et al., 2015). The fewer tests used to tap into specific constructs likely results from practical issues with data collection, where younger children have greater difficulty completing a longer battery of cognitive tests. However, this practical issue could explain why measurement models for younger samples tend to support unidimensional solutions: an insufficient number of construct-specific tests are administered, which limits the amount of construct-specific variance present in the model.

In terms of the consistency between adult and older adult studies, most older adult studies supported a two-factor solution, but there was also support for a three-factor solution. The three-factor models included inhibition, updating/working memory, and shifting, while the two-factor models either merged two of these factors or dropped one of them from the model. These findings could indicate a slight de-differentiation of abilities with older age; however, no studies supported a one-factor solution, a three-factor solution was supported in the oldest sample evaluated (Vaughan & Giovanello, 2010), and – unlike the preschool age group – all three factors were represented in at least one of the measurement models evaluated within this age band. As well, researchers have yet to evaluate the structure of executive function for a substantial portion of mid-life: none of the samples evaluated had a mean age between 30 and 60 years. In turn, if executive functions do de-differentiate, the representation of ages within the current review is not comprehensive enough to identify the time of life at which this de-differentiation occurs, indicating the need for more research on samples in middle adulthood along with more longitudinal investigations. The only longitudinal study evaluating changes in executive functions among older adults included in this review involved just two time points separated by a three-year interval among adults already aged 55 years and above (de Frias et al., 2009), which is an insufficient study duration to examine this issue. Overall, the results from the systematic review do not support the de-differentiation of executive functions with older age, with the caveat that there are insufficient longitudinal studies on the structure of executive functions and large gaps in the age spans represented in cross-sectional research.

The qualitative analysis effectively summarizes the previous latent variable research on the structure of executive functions, synthesizing the published findings that have followed the seminal work of Miyake and colleagues (2000). It is clear by the synthesis that the three factors evaluated by this original study (i.e., inhibition, updating/working memory, and shifting) have become the most frequently evaluated constructs within this field of research.

The extensive popularity of the three-factor model has offered a scaffold for the many reviews on executive function literature (e.g., Bardikoff & Sabbagh, 2017; Best & Miller, 2010; Best et al., 2009; Collette et al., 2006; Garon et al., 2008; Müller & Kerns, 2015; Niendam et al., 2012), where these three factors are often those most extensively discussed. The qualitative synthesis demonstrates that few researchers have expanded beyond the evaluation of these three factors, with few studies including other posited constructs (e.g., strategic retrieval, access to long-term memory) in their executive function measurement models (Adrover-Roig et al., 2012; Fournier-Vicente et al., 2008). Based on this research synthesis, there seems to be a general acceptance of the original three-factor measurement model (Miyake et al., 2000), with limited research pioneering beyond this set of factors throughout the lifespan. Many of these publications are conceptual replications, and their abundance may result from a publication bias in favor of a highly-cited model that many researchers have accepted as the standard model of the field. Despite these many conceptual replications, there is a merited concern about the replicability of this model, as made clear by the re-analysis.

Findings from the Re-Analysis

The re-analysis effort aimed to explore how well seven alternative models fit the data across multiple samples and test batteries. The re-analysis results were interpreted in two ways. First, by the rate of model acceptance, which considered the rate at which a model met conventional fit thresholds (i.e., CFI ≥ .90/.95; RMSEA ≤ .05/.08). This first method only evaluated the rate at which different models would converge and show acceptable fit among the 5,000 bootstrapped samples; it did not directly compare different models based on fit indices. The second interpretation was the rate of model selection. This method compared different models based on differences between their CFI and RMSEA values, determining which models presented with superior fit to other models. These results benefit from an interpretation in combination: model acceptance informs the rate at which a model fits the data, and model selection informs the rate at which a model has superior fit to an alternative model.

An important caveat regarding model selection is the calculation of two statistics: model selection and contingent model selection. Model selection quantifies the rate at which a model both converges and is selected over all other models among the 5,000 bootstrapped samples. If a model does not converge for a specific sample, it cannot be selected. In contrast, contingent model selection is the rate at which a model is selected over all other models among samples in which that model converges. Among samples where that model converges, the percentage quantifies the rate at which the model is superior to alternative models.

The most telling findings from this re-analysis was the remarkably low rate at which many published models converged and/or met fit thresholds among bootstrapped samples. Most of the studies included in the systematic review were of good quality (e.g., 80% of studies had a study quality score of ≥ 8/11), although very few had sufficient power (e.g., 20% $\hat{\pi} \geq .80$). The importance of statistical power in structural equation modeling has high relevance to the interpretation of these findings. Although rarely discussed by researchers publishing

executive function measurement models, the power of their models is contingent on sample size, model complexity, and the construct reliability of factors (Gagne & Hancock, 2006; Hancock, 2006; Wolf, Harrington, Clark, & Miller, 2013). Despite these issues, most studies included in the re-analysis had relatively small sample sizes and all tested a complex three-factor measurement model. Further, as observed in previous re-analyses of executive function measurement models, factors within this field often have weak to moderate levels of reliability, suggesting limited construct-specific variance captured by the latent factors (Willoughby et al., 2014). This low reliability results from low inter-item correlations between indicators, which was evident among studies included in the re-analysis (range of mean inter-item correlations: $r = 0.23$ to 0.39).

The low rate of model convergence may derive in part from the low construct reliability of factors included in the models, where a limited amount of true construct variance is present for the factors specified (Gagne & Hancock, 2006). In the current re-analysis, the models that converged the least often on average were those with the most factors. For example, the bifactor converged very rarely among adult samples, because there needed to be sufficient unique variance in the common factor, and all specific factors, to ensure adequate construct reliability and non-zero loadings. Alternatively, it is also possible that the low convergence rate resulted from highly similar loadings among indicators within the same factor, which has also been associated with issues of model identification (Kenny & Kashy, 1992). In the original selection of a nested factor model, the decision to drop the inhibition-specific factor was guided by low loadings onto this factor in the context of a bifactor model (Friedman et al., 2008). Considering the low reliability (Baggetta & Alexander, 2016; Schmidt, 2003) and low inter-test correlations often observed for executive function tests (Willoughby et al., 2014), the manifest variables included in the re-analysis could have had limited construct variance related to the factor(s) on which they loaded (Müller & Kerns, 2015). In turn, during the re-analysis effort, there may be insufficient construct-specific variance in the data for many of the models to properly converge.

A key question that can derive from these analyses is whether a lack of convergence is evidence against a true model. As articulated in the previous paragraph, multiple study-related design components can explain why a model does not properly converge, including sample size, model complexity, and the reliability of measurement. Although model complexity is associated with the study design, it is also associated with an underlying hypothesis about the structure of a construct. In the context of confirmatory factor analysis, study design intersects with the hypothesized structure of executive functions. This is a key reason why the rates of models meeting fit thresholds and contingent model selection were calculated, to determine the rates of model acceptance and selection regardless of convergence. However, considering the extremely low rates of convergence for some models, an interpretation of solely these values does not take all relevant information into consideration. For example, the bifactor model converges among only 24% of adult samples on average, but tends to fit more often than all other models among samples where it converges: 61 to 63% meet lenient RMSEA and CFI thresholds, respectively. The bifactor model also has a 49 to 55% rate of contingent model selection based on ΔCFI cutoffs. However, no published study has accepted the bifactor model, and its low rate of convergence undermines the support for this model based solely on evaluations of fit,

because it was not replicable in such a large proportion of bootstrapped samples. In turn, rates of convergence and fit have an interactive relationship, and the rates of model acceptance and selection offer the most effective method for summarizing this relationship: calculating the rate at which a model both properly converges and meets conventional fit thresholds or cutoffs.

A clear relationship existed between model complexity and convergence, in that more complex models converged less often. A relationship was also found between model complexity and model fit, where more complex models better fit the data. Low construct reliability may explain the high fit of complex models, where these models overfit the data and show excellent fit by explaining small amounts of covariation between tasks. When interpreting the re-analysis findings, these conflicting patterns made model selection a difficult task. While a unidimensional model almost always converges, it will almost never adequately fit the data among adults. In contrast, a nested factor model rarely converges, but when it does, it will more often meet traditional fit thresholds.

The excellent fit, low power, and poor construct reliability evident in published studies brings into question whether those models that fit well among a specific sample and specific battery of tests happen to be the models that get published, while other models that do not meet standard fit cutoffs remain in the file drawer. All published studies included in the qualitative synthesis reported excellent fit for their models (i.e., CFI ≥ .95; RMSEA ≤ .06), which provides no means for a reviewer of the overall literature to preferentially select one model from one study over an equally well-fitting model from another study. This concern aligns with the general concern of replicability currently facing psychological science (e.g., Pashler & Harris, 2012; Pashler & Wagenmakers, 2012; Simmons, Nelson, & Simonsoh, 2011).

A good fitting model captures the data well, but it does not necessarily reflect the true model for the population (Hancock, 2006). Considering the low power of these excellent fitting models, the question remains whether they could be replicated among small samples drawn from the same population. The majority of studies were underpowered and denoted as conceptual replications, rather than direct replications using identical test batteries and recruiting a sufficient sample size. These studies often found similar results to the first measurement model of executive functions (Miyake et al., 2000) despite using a different collection of tests and often an alternative population from which to sample. As with direct replication failures, conceptual replication failures are rarely published (Makel, Plucker, & Hegarty, 2012). In turn, it is possible that the many published studies that contain the most frequently reported factors (i.e., inhibition, updating, and shifting) may be the conceptual replication successes, while the failures not supporting a three-factor model remain in the file drawer.

One significant finding that may go missed from the aggregation of published work was that every published study found evidence for at least one measurement model. There were no studies that attempted to conceptually replicate a measurement model, failed, and published that failure. It is hard to imagine that a journal would eagerly publish a study involving solely a confirmatory factor analysis that did not report any model meeting standard fit

thresholds. Considering the heterogeneity of dependent variables across studies, researchers could adjust the indicators included in their model until they find a model that both converges and fits their data, either replacing or removing specific tests or re-analyzing the model with an alternative dependent variable for a given test. This approach would make the results of published studies highly data-driven; and explain, in part, the concerns of non-replicability deriving from the findings of the re-analysis.

A post-hoc analysis shed further light on the issue of publication bias and potential non-replicability within this field. On average, the accepted models reported by researchers were accepted among only around a third to less than half of bootstrapped child/adolescent (i.e., 33 to 36%) and adult samples (i.e., 37 to 44%) based on lenient fit thresholds. In terms of model selection, the originally selected model was only re-selected among about a third of child/adolescent samples (i.e., 31–37%) based on lenient cutoffs. The rates of re-selection were variable for adult samples depending on the use of ΔRMSEA (i.e., 14%) or ΔCFI (i.e., 38%). The rates of re-selection were only slightly higher based on contingent model selection, again using lenient cutoffs, among child/adolescent samples (i.e., 42% for both ΔRMSEA and ΔCFI) and adult samples (i.e., 22% using ΔRMSEA and 52% using ΔCFI).

These findings clearly illustrate a substantial publication bias across studies reporting measurement models for executive function. This bias affected the results of the re-analysis, which found low rates of model acceptance and selection for all the models evaluated, although some models appeared to fit the data or present with superior fit more consistently than others. Considering the influence of bias, the inference drawn from the re-analysis must be interpreted with significant caution. Issues of low power indicate that even the most established of models have weak evidence in aggregate. Further, publication bias may have resulted in the acceptance and dissemination of many studies that correspond to the widely accepted three-factor measurement model (Miyake et al., 2000). As articulated in the following section, the adult research does show modest support for the three-factor or nested factor models (e.g., Friedman et al., 2008, 2009, 2011, 2016), which could have resulted from researchers designing their studies around this model – which was apparent based on the qualitative synthesis – and reviewers preferring this model in their critique of submitted manuscripts. However, despite issues of publication bias, a primary aim of the re-analysis was to identify a measurement model that best fit the data across published studies; and the following interpretation of the re-analysis findings attempts to find a signal within the noise of re-analyzed data.

Re-analysis of adult samples—The published results offer some empirical information about the nature of executive functions. The statistician George Box once wrote “all models are wrong, but some are useful,” (Box & Draper, 1987, p. 424), which applies well to the current findings. As made visually clear by a peak in Figure 4c, the most frequently accepted factor model among adults was the nested factor model; however, this model only converged 57% of the time on average across samples. Among those samples for which the model converged, only 59% had an RMSEA ≤ 0.08 and only 54% had a CFI ≥ 0.90. In turn, despite being the most often accepted, the nested factor model would be accepted, based on lenient fit thresholds, among only 41 to 42% of 5,000 bootstrapped samples on average across studies. In regard to model selection, the nested factor model was the most often selected

based on ΔCFI lenient (i.e., 30%) and strict (i.e., 26%) cutoffs; however, based on ΔRMSEA, which penalizes for model complexity, the unidimensional model was selected most frequently per lenient (i.e., 26%) and strict (i.e., 32%) cutoffs. Although these peaks were present, per visual inspection of Figure 6a, they were not prominent, and alternative models (e.g., the three-factor model based on ΔCFI and the two-factor models based on ΔRMSEA) had similar rates of model selection. Presented as a peak in Figure 6b, the nested and bifactor models had the highest rates of contingent model selection based on ΔCFI; however, there was no clearly preferred model based on contingent model selection rates using ΔRMSEA, although the shifting-updating merged model was essentially never selected.

For the adult studies, three of the highest quality studies accepted the nested factor model using the same test battery across different samples (Fleming et al., 2016; Friedman et al., 2011; Ito et al., 2015). The results of these three studies align with the results of the overall re-analysis. The convergence rate ranged from 89% to 96% and the acceptance rate ranged from 72 to 96% and 83 to 95% for the lenient thresholds of the CFI and RMSEA, respectively. In terms of model selection, the nested factor model was selected among 39 to 70% and 10 to 22% of 5,000 bootstrapped samples based on the lenient cutoffs for the ΔCFI and ΔRMSEA, respectively. Among only those samples where the nested factor model converged, the rates of contingent model selection were largely similar: 41 to 76% and 11 to 24% of samples based on the lenient cutoffs for the ΔCFI and ΔRMSEA. Within a small set of consistent studies with well-powered, similarly aged samples ($\hat{\pi}$ range: 0.74 to 0.99; $\bar{x}$ age range: 17.30 to 22.50), the model consistently converged and met fit thresholds; however, it was inconsistently selected over alternative models.

The rates of model acceptance provide some support for the nested factor model among adult samples; however, when directly comparing different models based on changes in fit, no model was selected at a significantly greater frequency than other models among adults. When considering only those sample in which the nested factor model converges, the nested factor model was only selected at a higher rate based on ΔCFI cutoffs, while the use of ΔRMSEA cutoffs showed comparable rates of contingent model selection across most other models.

The RMSEA favors parsimonious models (Hooper et al., 2008), and the rates of model selection and contingent model selection based on ΔRMSEA indicate that more parsimonious models (e.g., unidimensional and two-factor models) tended to be selected more often than, or at similar rates to, the nested factor model. This finding could indicate that the nested factor model is too complex, with limited improvement in fit despite increased model complexity. However, both rates of model acceptance and model selection must be interpreted in combination. Whereas ΔRMSEA indicated the highest rate of model selection for the unidimensional model, this model was essentially never accepted based on conventional fit thresholds. As shown in Figure 4c, the nested factor model tended to be accepted most often based on lenient thresholds for both the CFI and RMSEA. In turn, even if the unidimensional model showed superior fit to a more complex model, it was extremely rare for this model to show acceptable fit, and it would not likely be accepted by a researcher evaluating competing models.

In aggregate, these results lend some tentative support for the nested factor model. This finding aligns with the basic premise of the first application of confirmatory factor analysis to executive functions (Miyake et al., 2000): the variance in executive function test batteries tends to show both unity and diversity. Although there is not a clear model that fully explains the precise structure of executive functions, the basic notion of unity and diversity is evident. A method for determining which measurement model ultimately aligns with the true nature of executive functions will require a closer examination of the brain-behavior relationships that underlie the constructs included in the accepted measurement model. Researchers have found brain activity during performance-based tasks of executive functions in areas associated with specific constructs, including the right inferior frontal cortex, basal ganglia, and pre-supplementary motor area activity during inhibition tasks (Aron, 2008), dorsolateral prefrontal cortex activity (DLPFC; Stuss & Levine, 2002) as well as frontopolar activity (Collette et al., 2005) during updating/working memory tasks, and DLPFC and dorsal anterior cingulate cortex activity during shifting tasks (Luna et al., 2015).

Although specific brain-behavior relationships have been proposed, there is evidence for both the unity and diversity of brain activity underlying separate executive function constructs (Collette et al., 2005, 2006). A comprehensive meta-analytic investigation (Niendam et al., 2012) found strong evidence for a superordinate fronto-cingulo-parietal network that showed common activity during tasks tapping into inhibition, working memory, and flexibility (i.e., a term often used synonymously with shifting; Baggetta & Alexander, 2016). This integrative function could parallel the common factor present in the nested factor model, which past researchers have conceptualized as the ability to “actively maintain task goals and goal-related information and use this information to effectively bias lower-level processing” (Miyake & Friedman, 2012, p. 11), arguably necessary for successful performance across executive function domains. Despite the alignment of the re-analysis findings and brain-behavior research, the results do not identify a definitive measurement model of executive function in adulthood. Considering issues of low power and publication bias, the findings are tentative, and require further scrutiny in future studies before any definitive model of executive function among adults can be unequivocally accepted.

Re-analysis of child/adolescent samples—In comparison to the findings among adult samples, the results of the re-analysis of the child/adolescent samples were interpretable in the opposite fashion. Whereas for the adult studies in Figure 4c, there was a clear peak in model acceptance rates for the three-factor and nested factors models, the child/adolescent studies in Figure 3c had two definitive “valleys” for the inhibition-updating merged and three-factor models, evidencing that models with differentiated shifting factors were *less* preferable to models that either merged the shifting factor or had a strong common executive function bifactor. This trend is consistent with discussion of a non-differentiated shifting factor early in development (Garon et al., 2008) and the notion that an independent shifting ability emerges later in development (Müller & Kerns, 2015). This trend was observed despite removing preschool samples from the means and medians calculated in the re-analysis.

The competing child/adolescent models that both converged and exceeded lenient fit thresholds most often were the unidimensional, shifting-updating merged, inhibition-shifting

merged, nested factor, and bifactor models. While these models were not easily differentiated based on the lenient CFI cutoff, the lenient RMSEA cutoff was met most often for the unidimensional ($\bar{x}$ = 32%; *Mdn* = 21%), the shifting-updating ($\bar{x}$ = 25%; *Mdn* = 32%), and inhibition-shifting models ($\bar{x}$ = 27%; *Mdn* = 30%). Considering the greater complexity of the nested factor and bifactor models, the more parsimonious models were favored by the RMSEA cutoff. As with the adult studies, there was not a clear determination about which model should be preferred based on fit indices; however, the re-analysis of child/adolescent samples supported (a) either a unidimensional or two-factor solution and (b) a model that does not have a differentiated shifting factor.

In comparison to the rates of model acceptance, the model selection analysis showed a clear peak in selection rates based on ΔRMSEA cutoffs in favor of the unidimensional model (i.e., 46 to 53% of samples), as shown visually in Figure 5a. Figure 5b shows this same peak for contingent model selection based on ΔRMSEA cutoffs (i.e., 46 to 54% of samples). Minimal differences in rates of model selection and contingent model selection were due to the mean 97% convergence rate of the unidimensional model. Contingent model selection did show a peak in favor of the nested factor model based on ΔCFI cutoffs (i.e. 27 to 30% of samples); however, this model was a distant second in rates of contingent selection based on ΔRMSEA cutoffs, and a more parsimonious interpretation would support a simpler unidimensional model. A comparison between Figures 3c and 5a showed the same pattern of valleys, where models with undifferentiated shifting factors tended to be selected at greater rates.

In combination, the results of the model acceptance and selection analyses lend the most support for a unidimensional model among the child/adolescent samples; however, this model was not accepted unequivocally, and two-factor models with an undifferentiated shifting factor had some modest levels of support as well. This non-differentiated system is supported by neurodevelopmental trajectories, where grey matter in the DLPFC, which is associated with both updating/working memory and shifting (Luna et al., 2015; Stuss & Levine, 2002), is pruned after the ventral frontal regions associated with inhibition (Aron, 2008) during child and adolescent development (Müller & Kerns, 2015). As with the adult findings, low power across these studies resulted in overall low rates of convergence and few models meeting traditional fit thresholds. In turn, these findings require a cautious interpretation; however, the conclusions are fairly conservative, and consistent with previous theories of executive function development (Bardikoff & Sabbagh, 2017; Garon et al., 2008).

Limitations

This systematic review and re-analysis offers the first comprehensive and empirical summary of measurement models for executive function test batteries across the lifespan. Despite the comprehensiveness of this review, the conclusions drawn from it remain tentative due to a variety of limitations. A first limitation pertains to the limited diversity of the samples evaluated. The eligible samples were largely balanced in gender (i.e., 52.56% female); however, the samples were not diverse in terms of their ethnic and racial composition. Ethnic or racial demographics were only reported for about 40% of samples, with clear discrepancies across age ranges in terms of how often this information was reported. Although 66% of preschool samples had racial or ethnic makeup reported, only

25% of older adult studies provided similar information. There were some studies with specifically Chinese samples (Duan et al., 2010; Lee et al., 2012; Xu et al., 2013) or majority minority samples (Masten et al., 2012; Rose et al., 2012); however, these few ethnically and racially diverse samples were exclusively child and adolescent.

Based on reported demographics, the adult and older adult samples were not only mostly White, but were also highly educated. Over half of the adult samples were undergraduate populations, while the older adults ranged in education from 11.30 to 17.67 years, with all but one sample having over 15 years of education on average. Based on the sample demographics, the generalizability of this research to diverse populations remains questionable. Furthermore, although the mean ages ranged from 3.01 to 73.68 across samples, there was still a gap in the representation of middle adulthood. As noted earlier, no researchers reported a sample with a mean age between 30 and 60. In turn, the structure of executive functions within middle adulthood remains largely unevaluated, because most studies categorized as adults in this review evaluated an undergraduate or college-aged sample. Future researchers would benefit from recruiting more participants within middle adulthood, without post-secondary education, and from diverse ethnic or racial backgrounds. This would ensure that the research findings on the structure of executive functions are representative beyond a well-educated and White population.

Additional limitations pertained specifically to the re-analysis effort. A primary aim of the re-analysis was to determine which published measurement model best fit summary data across studies; however, the results did not identify a best model, but rather showed modest levels of evidence for a small selection of models. Rates of model acceptance were overall quite low, even for the most often accepted model. Further, direct comparisons between models did not demonstrate a single model being accepted unequivocally. A reason for this finding may have resulted from the bootstrapping method, which cannot control for certain limitations of individual studies (e.g., low power, poor construct reliability). While a meta-analytic confirmatory factor analysis more effectively controls for these limitations when aggregating information across studies, this method relies on a pooled correlation matrix (Cheung & Chan, 2005), which requires the same variables to be used across different studies. Unfortunately, only a very small number of studies had the same set of manifest variables, thus impeding the use of a traditional meta-analytic approach. An assumption of confirmatory factor analysis is that the manifest variables are inter-changeable, which has led the field of executive function measurement models to include numerous different combinations of variables posited to tap into different constructs. The bootstrap method used here allowed for the synthesis of findings across studies using different test batteries, but every bootstrap iteration carries with it the individual limitations of the original empirical study. However, from another perspective, this apparent limitation did provide insight into the process of decision making at the modeling stage: simple models converge more often, but fail to fit the data well; while complex models hardly converge, but if they do, they tend to fit well. This resulted in just a small number of models that made it through the vetting process, and it explains the situation in this field, where a multitude of different factor structures tend to emerge, but each one of them is difficult to replicate.

Some analytical decisions and assumptions may also limit the interpretation of the current findings. In the re-analysis effort, residual correlations between tests were not specified, and no model modifications were considered. Such residual correlations or modifications could resolve issues of non-convergence or poor fit, and are often included for justifiable reasons in research practice. Conceptually, if each model for each of the 5,000 bootstrapped samples was closely examined, some model modifications could have allowed models to converge or improve fit; however, a model-by-model assessment at this level was not possible considering the magnitude of simulated samples and models evaluated, and this method could have resulted in the aggregation of fairly incomparable models depending on the extent of modifications needed for each model in individual samples. Another analytical decision that serves as a potential limitation was the wide age ranges used for both the child/adolescent ($\bar{x}$ age range: 8.33 to 14.41) and adult samples ($\bar{x}$ age range: 17.30 to 72.24). This decision limited inference about the structure of executive functions at specific points in human development (e.g., childhood vs. adolescence, young vs. older adulthood). Collapsing across developmental periods ensured a roughly equal number of samples fell within the child/adolescent ($k = 9$) and adult ($k = 10$) age spans prior to calculating a mean and median for rates of convergence and model acceptance or selection. Developmental considerations were taken prior to calculating means and medians during the re-analysis, such as excluding preschool samples due to a non-differentiated shifting factor (Miller et al., 2012; Usai et al., 2014). Despite wide age bands, conclusions based on a larger collection of samples arguably allow for more accurate inference about the structure of executive functions during development and adulthood.

Another limitation of the systematic review was the lack of individual participant data, because the findings presented in the re-analysis were based solely on simulated data using correlation matrices. Non-parametric bootstrapping with re-sampling is a more common method used by researchers with their raw datasets, but was not possible using summary data. If researchers were to use non-parametric bootstrapping with re-sampling to re-analyze their own sample data, the conclusions may differ from those amalgamated in the current review. In the context of the re-analysis, the parametric bootstrapping simulates samples of the same N as the observed samples, pulled from an assumed multivariate normal distribution. The alternative non-parametric bootstrapping with re-sampling approach more commonly used with raw data would not make this assumption; and software packages commonly used for confirmatory factor analysis would not offer a confidence interval around fit indices, nor a rate at which simulated samples met fit cutoffs. However, some software packages (e.g., MPlus; Muthén & Muthén, 2014) would quantify the number of bootstrapped draws completed, which would give an estimate of how often the model would properly converge. The use of bootstrapping may be fruitful for future researchers to guide their model selection, allowing them to determine the frequency at which an excellent fitting model would replicate among a set of bootstrapped samples.

Future Directions in Research on Executive Functions

In terms of future directions for researchers evaluating measurement models of executive functions, many gaps in the field remain unresolved based on the current review. As is clear from the findings, the results provided some guidance regarding which models have the most

– or least – empirical support, but they did not suggest that any model should be unequivocally accepted. Future researchers should evaluate alternative models including factors not previously represented in published measurement models. Despite some inconsistencies in the naming of factors, most researchers have taken the approach of evaluating the three-factor model (i.e., inhibition, updating, and shifting; Miyake et al., 2000), which has substantially influenced their test selection and design. The field of executive function measurement models shows a broad acceptance of the three-factor model, or the more recently proposed nested factor model of Miyake and Friedman (2012); however, the current findings raise serious doubts about the replicability of both of these models. Although there have been many conceptual or direct replications of these models (e.g., Lehto et al., 2003; de Frias et al., 2009; Fleming et al., 2016; Ito et al., 2015; Klauer et al., 2010), the re-analysis indicated only modest evidence for either of these models in aggregate. To move the field forward, researchers must continue to conduct high-powered studies to further evaluate and compare the replicability of these models, or include the assessment of new models or executive function factors not often evaluated by previous researchers.

Just a small set of studies explored additional constructs (e.g., Access to Long Term Memory, Adrover-Roig et al., 2012; Hot and Cool Executive Function, Carlson et al., 2014; Strategic Retrieval, Fournier-Vicente et al., 2008). Future researchers should consider exploring new constructs that have been postulated in previous research, but not consistently evaluated in confirmatory factor analyses, such as planning, problem solving, fluency, and reasoning (Packwood et al., 2011). As well, factor analytic studies not covered in this review have explored the multidimensionality of specific executive function constructs (e.g., inhibition, Aichert et al., 2012, Friedman & Miyake, 2004; problem solving, Cinan et al., 2013; Scherer & Tiemann, 2014), indicating that sub-components under the umbrella term of executive functions may be umbrella terms within themselves and worth further exploration.

In addition to the measurement of different constructs, other methods for advancing the field could include evaluating previously untested measurement models, re-analyzing primary datasets, or adding longitudinal follow-ups to research designs. Since the systematic search was conducted, one study evaluated a second-order factor model of executive functions (Wolff et al., 2016) and another tested a bifactor model that examined the differentiation of executive functions from fundamental cognitive abilities over the preschool years (Nelson et al., 2016). One recent re-analysis explored a formative factor model as an alternative method of both modeling and interpreting performances on tests of executive functions (Willoughby & Blair, 2016). While a formative model simply flips the directional path between manifest variables and factors (Kline, 2006), other re-analyses could conceptualize executive functions in a more causal manner. If conceptualizations of executive functions in early childhood suggest that inhibition and updating precede shifting development (Garon et al., 2008), then an alternative model could use causal paths, where shifting is endogenous to inhibition and updating in a structural equation modeling framework. In terms of longitudinal follow-up, only a small set of studies have evaluated longitudinal invariance of executive function factors (e.g., de Frias et al., 2009; Friedman et al., 2016; Lee et al., 2013;

Willoughby et al., 2012b), and future longitudinal research designs may clarify which factor structures are stable and replicable over time.

Future researchers would also benefit from conducting *a priori* power analyses before testing measurement models (Hancock, 2006), helping determine the necessary sample size to conduct their analysis. The systematic review clearly evidenced the issue of power endemic within this field, and future small-scale studies that do not consider power in their research design may ultimately be non-replicable. Any consumer of executive function research should be mindful of inferences drawn from underpowered studies with complex models explaining weak inter-item correlations, and future researchers within this field should explicitly address sample size, model complexity, and construct reliability as they relate to the power of their measurement model. This recommendation is not to dissuade researchers from conducting future confirmatory factor analyses on executive function test batteries, but rather to emphasize the importance of ensuring those future studies have the power to produce accurate and replicable findings.

When considering future small-scale studies, the consistency of the tests used to measure executive functions is of the utmost importance. The field must move towards a more consistent use of common tests with greater reliability to ensure that published measurement models are directly comparable and include factors with sufficient shared variance between manifest variables. While some tests were used consistently (e.g., the Stroop task, Antisaccade, *n*-back), a post-hoc exploration of the Stroop paradigms identified inconsistencies in the dependent variables that were derived from Stroop tests and ultimately used as indicators in measurement models. There were six different dependent variables deriving from Stroop paradigms among child/adolescent studies, as well as six different dependent variables deriving from Stroop paradigms among adult studies. Differences in the dependent variables deriving from specific tasks can potentially account for different results across studies. While this evaluation of dependent measures was a post-hoc exploration based on the published literature, it evidences the need for a close evaluation of the methods through which researchers measure executive functions in latent variable studies. An assumption of confirmatory factor analysis is that the manifest variables are interchangeable; however, different scores from the same test rarely correlate perfectly, and will have different relationships with other indicators and the latent factor. Thus, deciding on the tests used to measure specific constructs, and the scores used to operationalize these constructs, can have a substantial influence on the convergence and fit of a measurement model. The last review on the instruments used to assess executive functions occurred roughly a decade ago (Chan et al., 2008), and the current review provides a scaffold through which a closer examination of both executive function tests and scores can be evaluated. While the tests used by researchers vary by population (e.g., young children complete simpler paradigms than young adults), differences in the dependent variables deriving from these tests have not been explored. The post-hoc assessment of the Stroop test alluded to notable variability in the dependent measures used by different researchers examining different age groups. Conceptually, if researchers systematically differ in their preferred dependent variables (e.g., accuracy, reaction time, or a time-to-accuracy ratio), this methodological difference could explain some of the variability in the results observed across studies, and a closer

examination of heterogeneity in dependent measures moving forward could further the argument for greater consistency in executive function measurement.

Aside from variability in the exact scores used across confirmatory factor analyses, there was substantial variability in the batteries used across studies as well. Concerns about the heterogeneity between studies in how specific constructs are measured has been raised by previous reviewers of executive function research (Bardikoff & Sabbagh, 2017; Müller & Kerns, 2015). Although there is some consistency in the indicators assigned to different constructs, few studies had the exact same test battery, which could explain the inconsistencies in factor solutions and inter-factor correlations across different studies. Three of the highest quality studies were based on a common test battery (Fleming et al., 2016; Friedman et al., 2011; Ito et al., 2015), and all three accepted the nested factor model. The factor structure of this battery has also been evaluated longitudinally, showing stability in its structure over a 6-year period (Friedman et al., 2016).

The evaluation of executive functions in clinical practice is similarly disparate (Rabin et al., 2016). Since the first published measurement model on executive function, there has been a push for the translation of latent variable research into clinical practice (Miyake, Emerson, & Friedman, 2000), but practitioners do not often use composite scores of executive functions in their assessments. The continued evaluation of executive functions in both academic and clinical settings will require consistent measurement in order to provide comparable and interpretable results; however, any consensus in regards to its measurement would likely require an updated review of the many tests used to measure specific constructs to date (Chan et al., 2008), and a gathering of top researchers in the field to arrive at a preferred battery with a strong psychometric foundation to rationalize its widespread use (Baggetta & Alexander, 2016). The use of a common battery could overcome some of the shortcomings of individual studies evidenced by this review. A common battery would facilitate data sharing, and a data repository of common elements across studies would overcome issues of low power at the individual study level. Some researchers have attempted to produce batteries for widespread dissemination. The National Institute of Health funded the development of a test battery for the assessment of executive functions in clinical trials (i.e., Executive Abilities: Measures and Instruments for Neurobehavioral Evaluation and Research, EXAMINER; Kramer et al., 2014), providing factor scores for working memory, fluency, cognitive control, and a global composite, which align at least partly with the factors supported by the re-analysis of adult samples.

Conclusions

The systematic review and re-analysis offers the first comprehensive qualitative and quantitative synthesis of a rich body of latent variable research on executive function measurement models. This synthesis was conducted with three aims in mind: (a) summarizing the published evidence for different measurement models of executive functions, (b) identifying the number of executive function constructs evaluated as factors in previous studies, and (c) determining a best-fitting measurement model through re-analysis of summary data. The pursuit of these specific aims led to many relevant conclusions from a close evaluation of the published literature, as listed below:

- The constructs most often represented in published measurement models of executive function include inhibition, updating/working memory, and shifting.
- Published measurement models were most often one to two-factor models among preschoolers, three-factor models among school-aged children, three-factor or nested factor models among adolescents and adults, and two-factor models among older adults.
- These findings support differentiation of executive functions from preschool into adulthood, with the emergence of shifting during the school-age to adolescent years.
- The results do not offer support for the de-differentiation of executive functions over the course of adulthood, because the oldest sample evaluated produced a three-factor model and much of the adult age span (i.e., 30 to 60 years) is unrepresented in published research.
- For all models evaluated, the re-analysis showed predominantly low rates of model acceptance (i.e., the rate at which a model both converged and met conventional fit thresholds) and model selection (i.e., the rate at which a model converges and shows superior fit to all other models), which likely resulted from issues of low power and poor construct reliability when evaluating fairly complex measurement models.
- The re-analysis provided modest support for a one to two factor model among child/adolescent samples and a nested factor model among adult samples, which suggests greater unity among younger samples and a balance of unity and diversity among adult samples. However, considering low rates of model acceptance and selection overall, these findings are tentative, and no model was accepted unequivocally.
- Future researchers using confirmatory factor analysis should conduct *a priori* power analyses when designing their studies, considering sample size, model complexity, and construct reliability. Underpowered studies with complex models explaining limited shared variance will add non-replicable findings to the field.
- Moving forward, researchers should continue to determine the replicability of the models tested herein through high powered studies, but should also consider alternative models that may take a different approach to conceptualizing executive functions.

Overall these findings are tentative and do not offer definitive conclusions regarding the true nature of executive functions. Alternatively, the findings provided herein offer an affirmation of the “elusive nature of executive functions” (Jurado & Rosselli, 2007, p. 213). Despite its elusive nature, the goals of defining, measuring, and understanding executive functions remain tantamount to psychological research, considering the many clinical and functional outcomes associated with executive functions (e.g., Bell-McGinty et al., 2002; Cahn-Weiner et al., 2002; Espy et al., 2011; Karr, Areshenkoff, & Garcia-Barrera, 2014; Snyder, 2013; Scott et al., 2015) and the interventions already developed to enhance executive functions

across the lifespan (e.g., Baggetta & Alexander, 2016; Diamond & Lee, 2011; Karr, Areshenkoff, Rast, & Garcia-Barrera, 2014; Krasny-Pacini, Chevignard, & Evans, 2014).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Justin E. Karr is a Vanier Canada Graduate Scholar and thanks the Natural Sciences and Engineering Research Council of Canada (NSERC) for their support of his graduate studies. This study was completed in partial fulfillment of the requirements for his dissertation. Research reported in this publication was supported, in part, by NSERC Grant Number 418676-2012, Characteristics and Correlates of Intraindividual Variability in Executive Control Processes, to Mauricio A. Garcia-Barrera; National Institute on Aging of the National Institutes of Health under Award Number R01AG050720 to Philippe Rast; and Award Number P01AG043362 (2013–2018), Integrative Analysis of Longitudinal Studies of Aging and Dementia, to Scott M. Hofer. Grant L. Iverson acknowledges support from the United States Department of Defense as part of the TBI Endpoints Development Initiative with a grant entitled Development and Validation of a Cognition Endpoint for Traumatic Brain Injury Clinical Trials (subaward from W81XWH-14-2-0176). The content is solely the responsibility of the authors and does not necessarily represent the official views of NSERC, the National Institutes of Health, or the Department of Defense.

References

*Indicates a study included in the systematic review.

*. Adrover-Roig D, Sesé A, Barceló F, Palmer A. A latent variable approach to executive control in healthy ageing. Brain and Cognition. 2012; 78(3):284–299. DOI: 10.1016/j.bandc.2012.01.005 [PubMed: 22296984]

*. Agostino A, Johnson J, Pascual-Leone J. Executive functions underlying multiplicative reasoning: Problem type matters. Journal of Experimental Child Psychology. 2010; 105(4):286–305. DOI: 10.1016/j.jecp.2009.09.006 [PubMed: 19913238]

Aichert DS, Wöstmann NM, Costa A, Macare C, Wenig JR, Möller H, … Ettinger U. Associations between trait impulsivity and prepotent response inhibition. Journal of Clinical and Experimental Neuropsychology. 2012; 34(10):1016–1032. DOI: 10.1080/13803395.2012.706261 [PubMed: 22888795]

Allan NP, Lonigan CJ. Examining the dimensionality of effortful control in preschool children and its relation to academic and socioemotional indicators. Developmental Psychology. 2011; 47(4):905–915. DOI: 10.1037/a0023748 [PubMed: 21553957]

Allan NP, Lonigan CJ. Exploring dimensionality of effortful control using hot and cool tasks in a sample of preschool children. Journal of Experimental Child Psychology. 2014; 122:33–47. DOI: 10.1016/j.jecp.2013.11.013 [PubMed: 24518050]

Alvarez JA, Emory E. Executive function and the frontal lobes: A meta-analytic review. Neuropsychology Review. 2006; 16(1):17–42. DOI: 10.1007/s11065-006-9002-x [PubMed: 16794878]

Ardila A, Pineda DA. Factor structure of nonverbal cognition. International Journal of Neuroscience. 2000; 104(1):125–144. DOI: 10.3109/00207450009035013 [PubMed: 11011978]

*. Aran-Filippetti V. Structure and invariance of executive functioning tasks across socioeconomic status: evidence from spanish-speaking children. The Spanish Journal of Psychology. 2013; 16:1–15. DOI: 10.1017/sjp.2013.102

Aron AR. Progress in executive-function research: From tasks to functions to regions to networks. Current Directions in Psychological Science. 2008; 17(2):124–129.

Baddeley A, Hitch G. Working memory. In: Bower GH, editorRecent advances in learning and motivation. New York: Academic Press; 1974. 47–89.

Baena E, Allen PA, Kaut KP, Hall RJ. On age differences in prefrontal function: the importance of emotional/cognitive integration. Neuropsychologia. 2010; 48(1):319–333. DOI: 10.1016/j.neuropsychologia.2009.09.021 [PubMed: 19786039]

Baggetta P, Alexander PA. Conceptualization and operationalization of executive function. Mind, Brain, and Education. 2016; 10(1):10–33. DOI: 10.1111/mbe.12100

Barbey AK, Colom R, Grafman J. Dorsolateral prefrontal contributions to human intelligence. Neuropsychologia. 2013; 51(7):1361–1369. DOI: 10.1016/j.neuropsychologia.2012.05.017 [PubMed: 22634247]

Bardikoff N, Sabbagh M. The Differentiation of Executive Functioning Across Development: Insights from Developmental Cognitive Neuroscience. In: Budwig N, Turiel E, Zelazo P, editorsNew Perspectives on Human Development. Cambridge: Cambridge University Press; 2017. 27–46.

Barkley RA. The executive functions and self-regulation: An evolutionary neuropsychological perspective. Neuropsychology review. 2001; 11(1):1–29. DOI: 10.1023/A:1009085417776 [PubMed: 11392560]

Barkley RA. Executive functions: What they are, how they work, and why they evolved. New York: Guilford Press; 2012.

Barroso F. An approach to the study of attentional components in auditory tasks. Journal of Auditory Research. 1983; 23(3):157–180. [PubMed: 6680720]

Bell-McGinty S, Podell K, Franzen M, Baird AD, Williams MJ. Standard measures of executive function in predicting instrumental activities of daily living in older adults. International Journal of Geriatric Psychiatry. 2002; 17(9):828–834. DOI: 10.1002/gps.646 [PubMed: 12221656]

Benedek M, Jauk E, Sommer M, Arendasy M, Neubauer AC. Intelligence, creativity, and cognitive control: The common and differential involvement of executive functions in intelligence and creativity. Intelligence. 2014; 46:73–83. DOI: 10.1016/j.intell.2014.05.007 [PubMed: 25278640]

Bentler PM, Bonett DG. Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin. 1980; 88:588–606. DOI: 10.1037/0033-2909.88.3.588

Best JR, Miller PH. A developmental perspective on executive function. Child Development. 2010; 81(6):1641–1660. DOI: 10.1111/j.1467-8624.2010.01499.x [PubMed: 21077853]

Best JR, Miller PH, Jones LL. Executive functions after age 5: Changes and correlates. Developmental Review. 2009; 29(3):180–200. DOI: 10.1016/j.dr.2009.05.002 [PubMed: 20161467]

*. Bettcher BM, Mungas D, Patel N, Elofson J, Dutt S, Wynn M, … Kramer JH. Neuroanatomical substrates of executive functions: Beyond prefrontal structures. Neuropsychologia. 2016; 85:100–109. DOI: 10.1016/j.neuropsychologia.2016.03.001 [PubMed: 26948072]

Booth JN, Boyle JM, Kelly SW. Do tasks make a difference? Accounting for heterogeneity of performance of children with reading difficulties on tasks of executive function: Findings from a meta-analysis. British Journal of Developmental Psychology. 2010; 28(1):133–176. DOI: 10.1348/026151009X485432 [PubMed: 20306629]

Box GP, Draper NR. Empirical model-building and response surfaces. Oxford: John Wiley & Sons; 1987.

Brock LL, Rimm-Kaufman SE, Nathanson L, Grimm KJ. The contributions of 'hot'and 'cool'executive function to children's academic achievement, learning-related behaviors, and engagement in kindergarten. Early Childhood Research Quarterly. 2009; 24(3):337–349. DOI: 10.1016/j.ecresq.2009.06.001

Brocki KC, Bohlin G. Executive functions in children aged 6 to 13: A dimensional and developmental study. Developmental Neuropsychology. 2004; 26(2):571–593. DOI: 10.1207/s15326942dn2602_3 [PubMed: 15456685]

*. Brocki KC, Tillman C. Mental set shifting in childhood: The role of working memory and inhibitory control. Infant and Child Development. 2014; 23(6):588–604. DOI: 10.1002/icd.1871

Brookshire B, Levin HS, Song JX, Zhang L. Components of executive function in typically developing and head-injured children. Developmental Neuropsychology. 2004; 25(1&2):61–83. DOI: 10.1080/87565641.2004.9651922 [PubMed: 14984329]

Browne MW, Cudeck R. Alternative ways of assessing model fit. In: Bollen KA, Long JS, editorsTesting structural equation models. Newbury Park, CA: Sage; 1993. 136–162.

Brydges CR, Fox AM, Reid CL, Anderson M. Predictive validity of the N2 and P3 ERP components to executive functioning in children: A latent-variable analysis. Frontiers in Human Neuroscience. 2014; 8(80):1–10. DOI: 10.3389/fnhum.2014.00080 [PubMed: 24474914]

Brydges CR, Fox AM, Reid CL, Anderson M. The differentiation of executive functions in middle and late childhood: A longitudinal latent-variable analysis. Intelligence. 2014; 47:34–43. DOI: 10.1016/j.intell.2014.08.010

*. Brydges CR, Reid CL, Fox AM, Anderson M. A unitary executive function predicts intelligence in children. Intelligence. 2012; 40(5):458–469. DOI: 10.1016/j.intell.2012.05.006

Bull R, Espy KA, Wiebe SA, Sheffield TD, Nelson JM. Using confirmatory factor analysis to understand executive control in preschool children: Sources of variation in emergent mathematic achievement. Developmental Science. 2011; 14(4):679–692. DOI: 10.1111/j.1467-7687.2010.01012.x [PubMed: 21676089]

Burgess PW. Theory and methodology in executive function research. In: Rabbitt P, editorMethodology of frontal and executive function. Hove, UK: Psychology Press; 1997. 79–113.

Cahn-Weiner DA, Boyle PA, Malloy PF. Tests of executive function predict instrumental activities of daily living in community-dwelling older individuals. Applied Neuropsychology. 2002; 9(3):187–191. DOI: 10.1207/S15324826AN0903_8 [PubMed: 12584085]

Canivez GL. Construct validity of the WISC-IV with a referred sample: Direct versus indirect hierarchical structures. School Psychology Quarterly. 2014; 29(1):38–51. DOI: 10.1037/spq0000032 [PubMed: 23895320]

Canivez GL, Kush JC. WISC–IV and WAIS–IV structural validity: Alternate methods, alternate results. Commentary on Weiss et al (2013a) and Weiss et al. (2013b). Journal of Psychoeducational Assessment. 2013; 31(2):157–169. DOI: 10.1177/0734282913478036

*. Carlson SM, White RE, Davis-Unger AC. Evidence for a relation between executive function and pretense representation in preschool children. Cognitive Development. 2014; 29:1–16. DOI: 10.1016/j.cogdev.2013.09.001

Cassidy AR. Executive function and psychosocial adjustment in healthy children and adolescents: A latent variable modelling investigation. Child Neuropsychology. 2016; 22(3):292–317. DOI: 10.1080/09297049.2014.994484 [PubMed: 25569593]

Chan RC, Shum D, Toulopoulou T, Chen EY. Assessment of executive functions: Review of instruments and identification of critical issues. Archives of Clinical Neuropsychology. 2008; 23(2):201–216. DOI: 10.1016/j.acn.2007.08.010 [PubMed: 18096360]

Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural Equation Modeling. 2007; 14(3):464–504. DOI: 10.1080/10705510701301834

Chen FF, West SG, Sousa KH. A comparison of bifactor and second-order models of quality of life. Multivariate Behavioral Research. 2006; 41(2):189–225. DOI: 10.1207/s15327906mbr4102_5 [PubMed: 26782910]

Cheung MWL, Chan W. Meta-analytic structural equation modeling: A two-stage approach. Psychological Methods. 2005; 10(1):40–64. DOI: 10.1037/1082-989X.10.1.40 [PubMed: 15810868]

Chow M, Conway AR. The scope and control of attention: Sources of variance in working memory capacity. Memory & Cognition. 2015; 43(3):325–339. DOI: 10.3758/s13421-014-0496-9 [PubMed: 25604642]

Christopher ME, Miyake A, Keenan JM, Pennington B, DeFries JC, Wadsworth SJ, Willcutt E, Olson RK. Predicting word reading and comprehension with executive function and speed measures across development: a latent variable analysis. Journal of Experimental Psychology: General. 2012; 141(3):470–488. DOI: 10.1037/a0027375m [PubMed: 22352396]

Chuderski A, Taraday M, Nęcka E, Smoleń T. Storage capacity explains fluid intelligence but executive control does not. Intelligence. 2012; 40:278–295. DOI: 10.1016/j.intell.2012.02.010

Cinan S, Özen G, Hampshire A. Confirmatory factor analysis on separability of planning and insight constructs. Journal of Cognitive Psychology. 2013; 25(1):7–23. DOI: 10.1080/20445911.2012.729035

Cirino PT, Chapieski LM, Massman PJ. Card sorting performance and ADHD symptomatology in children and adolescents with Tourette syndrome. Journal of Clinical and Experimental

Neuropsychology. 2000; 22(2):245–256. DOI: 10.1076/1380-3395(200004)22:2;1-1;FT245 [PubMed: 10779838]

Cohen J. A power primer. Psychological bulletin. 1992; 112(1):155–159. DOI: 10.1037/0033-2909.112.1.155 [PubMed: 19565683]

Collette F, Hogge M, Salmon E, Van der Linden M. Exploration of the neural substrates of executive functioning by functional neuroimaging. Neuroscience. 2006; 139(1):209–221. DOI: 10.1016/j.neuroscience.2005.05.035 [PubMed: 16324796]

Collette F, van der Linden M, Laureys S, Delfiore G, Degueldre C, Luxen A, Salmon E. Exploring the unity and diversity of the neural substrates of executive functioning. Human Brain Mapping. 2005; 25(4):409–423. DOI: 10.1002/hbm.20118 [PubMed: 15852470]

Dang CP, Braeken J, Colom R, Ferrer E, Liu C. Why is working memory related to intelligence? Different contributions from storage and processing. Memory. 2014; 22(4):426–441. DOI: 10.1080/09658211.2013.797471 [PubMed: 23745736]

de Frias CM, Dixon RA, Strauss E. Structure of four executive functioning tests in healthy older adults. Neuropsychology. 2006; 20(2):206–214. DOI: 10.1037/0894-4105.20.2.206 [PubMed: 16594781]

*. de Frias CM, Dixon RA, Strauss E. Characterizing executive functioning in older special populations: From cognitively elite to cognitively impaired. Neuropsychology. 2009; 23(6):778–791. DOI: 10.1037/a0016743 [PubMed: 19899836]

Deckel AW, Hesselbrock V. Behavioral and Cognitive Measurements Predict Scores on the MAST: A 3-Year Prospective Study. Alcoholism: Clinical and Experimental Research. 1996; 20(7):1173–1178. DOI: 10.1111/j.1530-0277.1996.tb01107.x

Decker SL, Hill SK, Dean RS. Evidence of construct similarity in executive functions and fluid reasoning abilities. International Journal of Neuroscience. 2007; 117(6):735–748. DOI: 10.1080/00207450600910085 [PubMed: 17454241]

Della Sala S, Gray C, Spinnler H, Trivelli C. Frontal lobe functioning in man: the riddle revisited. Archives of Clinical Neuropsychology. 1998; 13(8):663–682. DOI: 10.1016/S0887-6177(97)00093-0 [PubMed: 14590627]

Diamond A, Lee K. Interventions shown to aid executive function development in children 4 to 12 years old. Science. 2011; 333(6045):959–964. DOI: 10.1126/science.1204529 [PubMed: 21852486]

*. Duan X, Wei S, Wang G, Shi J. The relationship between executive functions and intelligence on 11- to 12-year-old children. Psychological Test and Assessment Modeling. 2010; 52(4):419–431.

Duggan EC, Garcia-Barrera MA. Executive functioning and intelligence. In: Goldstein S, Naglieri JA, Princiotta D, editorsHandbook of intelligence: Evolutionary theory, historical perspective, and current concepts. New York: Springer Publishing Co; 2015. 435–458.

Edwards P, Clarke M, DiGuiseppi C, Pratap S, Roberts I, Wentz R. Identification of randomized controlled trials in systematic reviews: Accuracy and reliability of screening records. Statistics in Medicine. 2002; 21(11):1635–1640. DOI: 10.1002/sim.1190 [PubMed: 12111924]

Egeland J, Landrø NI, Tjemsland E, Walbækken K. Norwegian norms and factor-structure of phonemic and semantic word list generation. The Clinical Neuropsychologist. 2006; 20(4):716–728. DOI: 10.1080/13854040500351008 [PubMed: 16980257]

Engel de Abreu PM, Gathercole SE. Executive and phonological processes in second-language acquisition. Journal of Educational Psychology. 2012; 104(4):974–986. DOI: 10.1037/a0028390

Engle RW, Tuholski SW, Laughlin JE, Conway AR. Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. Journal of Experimental Psychology: General. 1999; 128(3):309–331. DOI: 10.1037/0096-3445.128.3.309 [PubMed: 10513398]

Eslinger PJ. Conceptualizing, describing, and measuring components of executive function. In: Lyons GR, Krasnegor NA, editorsAttention, memory and executive function. Baltimore, MD: Brooks Publishing; 1996. 367–395.

Espy KA, Kaufmann PM, Glisky ML, McDiarmid MD. New procedures to assess executive functions in preschool children. The Clinical Neuropsychologist. 2001; 15(1):46–58. DOI: 10.1076/clin.15.1.46.1908 [PubMed: 11778578]

Espy KA, McDiarmid MM, Cwik MF, Stalets MM, Hamby A, Senn TE. The contribution of executive functions to emergent mathematic skills in preschool children. Developmental Neuropsychology. 2004; 26(1):465–486. DOI: 10.1207/s15326942dn2601_6 [PubMed: 15276905]

Espy KA, Sheffield TD, Wiebe SA, Clark CA, Moehr MJ. Executive control and dimensions of problem behaviors in preschool children. Journal of Child Psychology and Psychiatry. 2011; 52(1): 33–46. DOI: 10.1111/j.1469-7610.2010.02265.x [PubMed: 20500238]

Ettenhofer ML, Hambrick DZ, Abeles N. Reliability and stability of executive functioning in older adults. Neuropsychology. 2006; 20(5):607–613. DOI: 10.1037/0894-4105.20.5.607 [PubMed: 16938023]

Fan X, Thompson B, Wang L. Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. Structural Equation Modeling. 1999; 6(1):56–83. DOI: 10.1080/10705519909540119

Fisk JE, Sharp CA. Age-related impairment in executive functioning: Updating, inhibition, shifting, and access. Journal of Clinical and Experimental Neuropsychology. 2004; 26(7):874–890. DOI: 10.1080/13803390490510680 [PubMed: 15742539]

*. Fleming KA, Heintzelman SJ, Bartholow BD. Specifying associations between conscientiousness and executive functioning: Mental set shifting, not prepotent response inhibition or working memory updating. Journal of Personality. 2016; 84(3):348–360. DOI: 10.1111/jopy.12163 [PubMed: 25564728]

Floyd RG, Bergeron R, Hamilton G, Parra GR. How do executive functions fit with the Cattell–Horn–Carroll model? Some evidence from a joint factor analysis of the Delis–Kaplan executive function system and the Woodcock–Johnson III tests of cognitive abilities. Psychology in the Schools. 2010; 47(7):721–738. DOI: 10.1002/pits.20500

*. Fournier-Vicente S, Larigauderie P, Gaonac'h D. More dissociations and interactions within central executive functioning: A comprehensive latent-variable analysis. Acta Psychologica. 2008; 129(1):32–48. DOI: 10.1016/j.actpsy.2008.04.004 [PubMed: 18499078]

*. Frazier DT, Bettcher BM, Dutt S, Patel N, Mungas D, Miller J, … Kramer JH. Relationship between insulin-resistance processing speed and specific executive function profiles in neurologically intact older adults. Journal of the International Neuropsychological Society. 2015; 21(8):622–628. DOI: 10.1017/S1355617715000624 [PubMed: 26272269]

Friedman NP, Corley RP, Hewitt JK, Wright KP. Individual differences in childhood sleep problems predict later cognitive executive control. Sleep. 2009; 32(3):323–333. DOI: 10.1093/sleep/32.3.323 [PubMed: 19294952]

Friedman NP, Haberstick BC, Willcutt EG, Miyake A, Young SE, Corley RP, Hewitt JK. Greater attention problems during childhood predict poorer executive functioning in late adolescence. Psychological Science. 2007; 18(10):893–900. DOI: 10.1111/j.1467-9280.2007.01997.x [PubMed: 17894607]

Friedman NP, Miyake A. The relations among inhibition and interference control functions: A latent-variable analysis. Journal of Experimental Psychology: General. 2004; 133(1):101–135. DOI: 10.1037/0096-3445.133.1.101 [PubMed: 14979754]

Friedman NP, Miyake A, Altamirano LJ, Corley RP, Young SE, Rhea SA, Hewitt JK. Stability and change in executive function abilities from late adolescence to early adulthood: A longitudinal twin study. Developmental Psychology. 2016; 52(2):326–340. DOI: 10.1037/dev0000075 [PubMed: 26619323]

Friedman NP, Miyake A, Corley RP, Young SE, DeFries JC, Hewitt JK. Not all executive functions are related to intelligence. Psychological Science. 2006; 17(2):172–179. DOI: 10.1111/j.1467-9280.2006.01681.x [PubMed: 16466426]

*. Friedman NP, Miyake A, Robinson JL, Hewitt JK. Developmental trajectories in toddlers' self-restraint predict individual differences in executive functions 14 years later: A behavioral genetic analysis. Developmental Psychology. 2011; 47(5):1410–1430. DOI: 10.1037/a0023750 [PubMed: 21668099]

Friedman NP, Miyake A, Young SE, DeFries JC, Corley RP, Hewitt JK. Individual differences in executive functions are almost entirely genetic in origin. Journal of Experimental Psychology: General. 2008; 137(2):201–225. DOI: 10.1037/0096-3445.137.2.201 [PubMed: 18473654]

Fuhs MW, Day JD. Verbal ability and executive functioning development in preschoolers at head start. Developmental Psychology. 2011; 47(2):404–416. DOI: 10.1037/a0021065 [PubMed: 21142363]

Gagne P, Hancock GR. Measurement model quality, sample size, and solution propriety in confirmatory factor models. Multivariate Behavioral Research. 2006; 41(1):65–83. DOI: 10.1207/s15327906mbr4101_5 [PubMed: 26788895]

Gansler DA, Suvak M, Arean P, Alexopoulos GS. Role of executive dysfunction and dysexecutive behavior in late-life depression and disability. The American Journal of Geriatric Psychiatry. 2015; 23(10):1038–1045. DOI: 10.1016/j.jagp.2015.05.003 [PubMed: 26209224]

Garon N, Bryson SE, Smith IM. Executive function in preschoolers: A review using an integrative framework. Psychological Bulletin. 2008; 134(1):31–60. DOI: 10.1037/0033-2909.134.1.31 [PubMed: 18193994]

Garrett HE. A developmental theory of intelligence. American Psychologist. 1946; 1(9):372–378. DOI: 10.1037/h0056380 [PubMed: 20280373]

Garza JP, Nelson JM, Sheffield TD, Choi H, Clark CA, Wiebe SA, Espy KA. Parsing executive control from foundational cognitive abilities in preschool: Application of the bifactor model to examine developmental change. In: Espy KA, Chair, editorThe changing nature of executive control in preschool: Using statistical modeling to situate neuroscience in development; Symposium conducted at the Forty-Second Annual Meeting of the International Neuropsychological Society; Seattle, Washington. 2014 Feb.

Gay P, Rochat L, Billieux J, d'Acremont M, Van der Linden M. Heterogeneous inhibition processes involved in different facets of self-reported impulsivity: Evidence from a community sample. Acta Psychologica. 2008; 129(3):332–339. DOI: 10.1016/j.actpsy.2008.08.010 [PubMed: 18851842]

Goschke T. Dysfunctions of decision- making and cognitive control as transdiagnostic mechanisms of mental disorders: advances, gaps, and needs in current research. International Journal of Methods in Psychiatric Research. 2014; 23(S1):41–57. DOI: 10.1002/mpr.1410 [PubMed: 24375535]

Greiff S, Wüstenberg S, Funke J. Dynamic problem solving: A new assessment perspective. Applied Psychological Measurement. 2012; 36(3):189–213. DOI: 10.1177/0146621612439620

Greve KW, Stickle TR, Love JM, Bianchini KJ, Stanford MS. Latent structure of the Wisconsin Card Sorting Test: A confirmatory factor analytic study. Archives of Clinical Neuropsychology. 2005; 20(3):355–364. DOI: 10.1016/j.acn.2004.09.004 [PubMed: 15797171]

Grodzinsky GM, Diamond R. Frontal lobe functioning in boys with attention-deficit hyperactivity disorder. Developmental Neuropsychology. 1992; 8(4):427–445. DOI: 10.1080/87565649209540536

Gustafsson JE, Wolff U. Measuring fluid intelligence at age four. Intelligence. 2015; 50:175–185. DOI: 10.1016/j.intell.2015.04.008

Hancock GR. Power analysis in covariance structure modeling. In: Hancock GR, Mueller RO, editorsStructural equation modeling: A second course. Greenwich, Connecticut: Information Age Publishing; 2006. 69–115.

Haring L, Mõttus R, Koch K, Trei M, Maron E. Factorial validity, measurement equivalence and cognitive performance of the Cambridge Neuropsychological Test Automated Battery (CANTAB) between patients with first-episode psychosis and healthy volunteers. Psychological Medicine. 2015; 45(09):1919–1929. DOI: 10.1017/S0033291714003018 [PubMed: 25544472]

Harvey DJ, Beckett LA, Mungas DM. Multivariate modeling of two associated cognitive outcomes in a longitudinal study. Journal of Alzheimer's Disease. 2003; 5(5):357–365.

*. Hedden T, Yoon C. Individual differences in executive processing predict susceptibility to interference in verbal working memory. Neuropsychology. 2006; 20(5):511–528. DOI: 10.1037/0894-4105.20.5.511 [PubMed: 16938014]

Hooper D, Coughlan J, Mullen M. Structural equation modelling: Guidelines for determining model fit. Electronic Journal of Business Research Methods. 2008; 6(1):53–60.

Howard SJ, Okely AD. Catching fish and avoiding sharks: Investigating factors that influence developmentally appropriate measurement of preschoolers' inhibitory control. Journal of Psychoeducational Assessment. 2015; 33(6):585–596. DOI: 10.1177/0734282914562933 [PubMed: 26339119]

Howard SJ, Johnson J, Pascual-Leone J. Clarifying inhibitory control: Diversity and development of attentional inhibition. Cognitive Development. 2014; 31:1–21. DOI: 10.1016/j.cogdev.2014.03.001

Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling. 1999; 6:1–55. DOI: 10.1080/10705519909540118

*. Huizinga M, Dolan CV, Van der Molen MW. Age-related change in executive function: Developmental trends and a latent variable analysis. Neuropsychologia. 2006; 44(11):2017–2036. DOI: 10.1016/j.neuropsychologia.2006.01.010 [PubMed: 16527316]

Hughes C, Ensor R. Individual differences in growth in executive function across the transition to school predict externalizing and internalizing behaviors and self-perceived academic success at 6 years of age. Journal of Experimental Child Psychology. 2011; 108(3):663–676. DOI: 10.1016/j.jecp.2010.06.005 [PubMed: 20673580]

Hughes C, Graham A. Measuring executive functions in childhood: Problems and solutions. Child and Adolescent Mental Health. 2002; 7(3):131–142. DOI: 10.1111/1475-3588.00024

Hughes C, Ensor R, Wilson A, Graham A. Tracking executive function across the transition to school: A latent variable approach. Developmental Neuropsychology. 2010; 35(1):20–36. DOI: 10.1080/87565640903325691 [PubMed: 20390590]

*. Hull R, Martin RC, Beier ME, Lane D, Hamilton AC. Executive function in older adults: a structural equation modeling approach. Neuropsychology. 2008; 22(4):508–522. DOI: 10.1037/0894-4105.22.4.508 [PubMed: 18590362]

*. Ito TA, Friedman NP, Bartholow BD, Correll J, Loersch C, Altamirano LJ, Miyake A. Toward a comprehensive understanding of executive cognitive function in implicit racial bias. Journal of Personality and Social Psychology. 2015; 108(2):187–218. DOI: 10.1037/a0038557 [PubMed: 25603372]

Johnson MH. Functional brain development in infants: elements of an interactive specialization framework. Child Development. 2000; 71(1):75–81. DOI: 10.1111/1467-8624.00120 [PubMed: 10836560]

Johnson MH. Interactive specialization: a domain-general framework for human functional brain development? Developmental Cognitive Neuroscience. 2011; 1(1):7–21. DOI: 10.1016/j.dcn.2010.07.003 [PubMed: 22436416]

Jurado MB, Rosselli M. The elusive nature of executive functions: a review of our current understanding. Neuropsychology Review. 2007; 17(3):213–233. DOI: 10.1007/s11065-007-9040-z [PubMed: 17786559]

Karasinski C. Language ability, executive functioning and behaviour in school- age children. International Journal of Language & Communication Disorders. 2015; 50(2):144–150. DOI: 10.1111/1460-6984.12104 [PubMed: 25582151]

Karr JE, Areshenkoff CN, Garcia-Barrera MA. The neuropsychological outcomes of concussion: A systematic review of meta-analyses on the cognitive sequelae of mild traumatic brain injury. Neuropsychology. 2014; 28(3):321–336. DOI: 10.1037/neu0000037 [PubMed: 24219611]

Karr JE, Areshenkoff CN, Rast P, Garcia-Barrera MA. An empirical comparison of the therapeutic benefits of physical exercise and cognitive training on the executive functions of older adults: A meta-analysis of controlled trials. Neuropsychology. 2014; 28(6):829–845. DOI: 10.1037/neu0000101 [PubMed: 24933486]

Kenny DA, Kashy DA. Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. Psychological Bulletin. 1992; 112:165–172. DOI: 10.1037/0033-2909.112.1.165

Keren-Happuch E, Chen SA, Ho MR, Desmond JE. A meta-analysis of cerebellar contributions to higher cognition from PET and fMRI studies. Human Brain Mapping. 2014; 35(2):593–615. DOI: 10.1002/hbm.22194 [PubMed: 23125108]

*. Klauer KC, Schmitz F, Teige-Mocigemba S, Voss A. Understanding the role of executive control in the Implicit Association Test: Why flexible people have small IAT effects. The Quarterly Journal of Experimental Psychology. 2010; 63(3):595–619. DOI: 10.1080/17470210903076826 [PubMed: 19672797]

Klenberg L, Korkman M, Lahti-Nuuttila P. Differential development of attention and executive functions in 3-to 12-year-old Finnish children. Developmental Neuropsychology. 2001; 20(1): 407–428. DOI: 10.1207/S15326942DN2001_6 [PubMed: 11827096]

Kline RB. Reverse arrow dynamics: Formative measurement and feedback loops. In: Hancock GR, Mueller RO, editorsStructural equation modeling: A second course. Greenwich, Connecticut: Information Age Publishing; 2006. 43–68.

Kramer JH, Mungas D, Possin KL, Rankin KP, Boxer AL, Rosen HJ, … Widmeyer M. NIH EXAMINER: Conceptualization and development of an executive function battery. Journal of the International Neuropsychological Society. 2014; 20(1):11–19. DOI: 10.1017/S1355617713001094 [PubMed: 24103232]

Krasny-Pacini A, Chevignard M, Evans J. Goal Management Training for rehabilitation of executive functions: A systematic review of effectiveness in patients with acquired brain injury. Disability and Rehabilitation. 2014; 36(2):105–116. DOI: 10.3109/09638288.2013.777807 [PubMed: 23597002]

*. Lambek R, Shevlin M. Working memory and response inhibition in children and adolescents: Age and organization issues. Scandinavian Journal of Psychology. 2011; 52(5):427–432. DOI: 10.1111/j.1467-9450.2011.00899.x [PubMed: 21722136]

Latzman RD, Markon KE. The factor structure and age-related factorial invariance of the Delis-Kaplan Executive Function System (D-KEFS). Assessment. 2010; 17(2):172–184. DOI: 10.1177/1073191109356254 [PubMed: 20040723]

*. Lee K, Bull R, Ho RM. Developmental changes in executive functioning. Child Development. 2013; 84(6):1933–1953. DOI: 10.1111/cdev.12096 [PubMed: 23550969]

*. Lee K, Ng SF, Pe ML, Ang SY, Hasshim MNAM, Bull R. The cognitive underpinnings of emerging mathematical skills: Executive functioning, patterns, numeracy, and arithmetic. British Journal of Educational Psychology. 2012; 82(1):82–99. DOI: 10.1111/j.2044-8279.2010.02016.x [PubMed: 22429059]

Lee CS, Therriault DJ. The cognitive underpinnings of creative thought: A latent variable analysis exploring the roles of intelligence and working memory in three creative thinking processes. Intelligence. 2013; 41(5):306–320. DOI: 10.1016/j.intell.2013.04.008

*. Lehto JE, Juujärvi P, Kooistra L, Pulkkinen L. Dimensions of executive functioning: Evidence from children. British Journal of Developmental Psychology. 2003; 21(1):59–80. DOI: 10.1348/026151003321164627

*. Lerner MD, Lonigan CJ. Executive function among preschool children: Unitary versus distinct abilities. Journal of Psychopathology and Behavioral Assessment. 2014; 36(4):626–639. DOI: 10.1007/s10862-014-9424-3 [PubMed: 25642020]

Levin HS, Fletcher JM, Kufera JA, Harward H, Lilly MA, Mendelsohn D, … Eisenberg HM. Dimensions of cognition measured by the tower of London and other cognitive tasks in head-injured children and adolescents. Developmental Neuropsychology. 1996; 12(1):17–34. DOI: 10.1080/87565649609540638

Lezak MD. Neuropsychological assessment. New York: Oxford University Press; 1976.

Lezak MD. The problem of assessing executive functions. International Journal of Psychology. 1982; 17(1–4):281–297. DOI: 10.1080/00207598208247445

Lezak MD. Neuropsychological assessment. 2. New York: Oxford University Press; 1983.

Lin F, Roiland R, Chen DGD, Qiu C. Linking cognition and frailty in middle and old age: Metabolic syndrome matters. International Journal of Geriatric Psychiatry. 2015; 30(1):64–71. DOI: 10.1002/gps.4115 [PubMed: 24733716]

Luna B, Marek S, Larsen B, Tervo-Clemmens B, Chahal R. An integrative model of the maturation of cognitive control. Annual Review of Neuroscience. 2015; 38:151–170. DOI: 10.1146/annurev-neuro-071714-034054

Luria AR. Higher cortical functions in man. New York: Basic Books; 1966.

Makel MC, Plucker JA, Hegarty B. Replications in psychology research how often do they really occur? Perspectives on Psychological Science. 2012; 7(6):537–542. DOI: 10.1177/1745691612460688 [PubMed: 26168110]

Mann SL, Selby EA, Bates ME, Contrada RJ. Integrating affective and cognitive correlates of heart rate variability: A structural equation modeling approach. International Journal of Psychophysiology. 2015; 98(1):76–86. DOI: 10.1016/j.ijpsycho.2015.07.003 [PubMed: 26168884]

Marcovitch S, O'Brien M, Calkins SD, Leerkes EM, Weaver JM, Levine DW. A longitudinal assessment of the relation between executive function and theory of mind at 3, 4, and 5 years. Cognitive Development. 2015; 33:40–55. DOI: 10.1016/j.cogdev.2014.07.001 [PubMed: 25642021]

Marcus GF, Davis E. How robust are probabilistic models of higher-level cognition? Psychological Science. 2013; 24(12):2351–2360. DOI: 10.1177/0956797613495418 [PubMed: 24084039]

*. Masten AS, Herbers JE, Desjardins CD, Cutuli JJ, McCormick CM, Sapienza JK, … Zelazo PD. Executive function skills and school success in young children experiencing homelessness. Educational Researcher. 2012; 41(9):375–384. DOI: 10.3102/0013189X12459883

McAuley T, White DA. A latent variables examination of processing speed, response inhibition, and working memory during typical development. Journal of Experimental Child Psychology. 2011; 108(3):453–468. DOI: 10.1016/j.jecp.2010.08.009 [PubMed: 20888572]

McFall GP, Wiebe SA, Vergote D, Jhamandas J, Westaway D, Dixon RA. IDE (rs6583817) polymorphism and pulse pressure are independently and interactively associated with level and change in executive function in older adults. Psychology And Aging. 2014; 29(2):418–430. DOI: 10.1037/a0034656 [PubMed: 24660790]

McFall GP, Wiebe SA, Vergote D, Westaway D, Jhamandas J, Dixon RA. IDE (rs6583817) polymorphism and type 2 diabetes differentially modify executive function in older adults. Neurobiology Of Aging. 2013; 34(9):2208–2216. DOI: 10.1016/j.neurobiolaging.2013.03.010 [PubMed: 23597493]

McVay JC, Kane MJ. Why does working memory capacity predict variation in reading comprehension? On the influence of mind wandering and executive attention. Journal of Experimental Psychology: General. 2012; 141(2):302–320. DOI: 10.1037/a0025250 [PubMed: 21875246]

*. Miller MR, Giesbrecht GF, Müller U, McInerney RJ, Kerns KA. A latent variable approach to determining the structure of executive function in preschool children. Journal of Cognition and Development. 2012; 13(3):395–423. DOI: 10.1080/15248372.2011.585478

Miller MR, Müller U, Giesbrecht GF, Carpendale JI, Kerns KA. The contribution of executive function and social understanding to preschoolers' letter and math skills. Cognitive Development. 2013; 28(4):331–349. DOI: 10.1016/j.cogdev.2012.10.005

Mitchell MB, Shaughnessy LW, Shirk SD, Yang FM, Atri A. Neuropsychological test performance and cognitive reserve in healthy aging and the Alzheimer's disease spectrum: A theoretically driven factor analysis. Journal of the International Neuropsychological Society. 2012; 18(06):1071–1080. DOI: 10.1017/S1355617712000859 [PubMed: 23039909]

Miyake A, Friedman NP. The nature and organization of individual differences in executive functions four general conclusions. Current Directions in Psychological Science. 2012; 21(1):8–14. DOI: 10.1177/0963721411429458 [PubMed: 22773897]

*. Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, Wager TD. The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: A latent variable analysis. Cognitive Psychology. 2000; 41(1):49–100. DOI: 10.1006/cogp.1999.0734 [PubMed: 10945922]

Miyake A, Emerson MJ, Friedman NP. Assessment of executive functions in clinical settings: Problems and recommendations. Seminars in Speech and Language. 2000; 21(2):169–183. DOI: 10.1055/s-2000-7563 [PubMed: 10879548]

Miyake A, Friedman NP, Rettinger DA, Shah P, Hegarty M. How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. Journal of Experimental Psychology: General. 2001; 130(4):621–640. DOI: 10.1037//0096-3445.130.4.621 [PubMed: 11757872]

Moher D, Liberati A, Tetzlaff J, Altman D. the PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. Annals of Internal Medicine. 2009; 151:264–269. DOI: 10.7326/0003-4819-151-4-200908180-00135 [PubMed: 19622511]

*. Monette S, Bigras M, Lafrenière MA. Structure of executive functions in typically developing kindergarteners. Journal of Experimental Child Psychology. 2015; 140:120–139. DOI: 10.1016/j.jecp.2015.07.005 [PubMed: 26241760]

Müller U, Kerns K. The development of executive function. In: Liben LS, Müller U, Lerner RM, editorsHandbook of child psychology and developmental science, Vol. 2: Cognitive processes. 7. Hoboken, NJ: Wiley; 2015. 571–623.

Muthén L, Muthén B. MPlus (Version 7.3) [Computer Software]. Los Angeles: Muthén & Muthén; 2014.

Nelson JM, James TD, Choi HJ, Clark CAC, Wiebe SA, Espy KA. Distinguishing executive control from overlapping foundational cognitive abilities during the preschool period. Monographs of the Society for Research in Child Development. 2016; 81(4):47–68. DOI: 10.1111/mono.12270 [PubMed: 27943324]

Niendam TA, Laird AR, Ray KL, Dean YM, Glahn DC, Carter CS. Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. Cognitive, Affective, & Behavioral Neuroscience. 2012; 12(2):241–268. DOI: 10.3758/s13415-011-0083-5

Nieuwenhuis S, Monsell S. Residual costs in task switching: Testing the failure-to-engage hypothesis. Psychonomic Bulletin & Review. 2002; 9(1):86–92. DOI: 10.3758/BF03196259 [PubMed: 12026956]

Norman DA, Shallice T. Attention to action: Willed and automatic control of behavior. In: Davidson RJ, Schwartz GE, Shapiro D, editorsConsciousness and self-regulation: Advances in research and theory. New York: Plenum; 1986. 1–18.

OCEBM (Oxford Centre for Evidence-Based Medicine) Levels of Evidence Working Group. The Oxford Levels of Evidence 2. 2011. Retrieved from http://www.cebm.net/index.aspx?o=5653

Oosterlaan J, Scheres A, Sergeant JA. Which executive functioning deficits are associated with AD/HD, ODD/CD and comorbid AD/HD+ ODD/CD? Journal of Abnormal Child Psychology. 2005; 33(1):69–85. DOI: 10.1007/s10802-005-0935-y [PubMed: 15759592]

Packwood S, Hodgetts HM, Tremblay S. A multiperspective approach to the conceptualization of executive functions. Journal of Clinical and Experimental Neuropsychology. 2011; 33(4):456–470. DOI: 10.1080/13803395.2010.533157 [PubMed: 21271425]

Pashler H, Harris CR. Is the replicability crisis overblown? Three arguments examined. Perspectives on Psychological Science. 2012; 7(6):531–536. DOI: 10.1177/1745691612463401 [PubMed: 26168109]

Pashler H, Wagenmakers EJ. Editors' introduction to the special section on replicability in psychological science a crisis of confidence? Perspectives on Psychological Science. 2012; 7(6):528–530. DOI: 10.1177/1745691612465253 [PubMed: 26168108]

*. Pettigrew C, Martin RC. Cognitive declines in healthy aging: Evidence from multiple aspects of interference resolution. Psychology and Aging. 2014; 29(2):187–204. DOI: 10.1037/a0036085 [PubMed: 24955989]

Phillips LH. Do "frontal tests" measure executive function? Issues of assessment and evidence from fluency tests. In: Rabbitt P, editorMethodology of frontal and executive function. Hove, UK: Psychology Press; 1997. 191–213.

Pickens S, Ostwald SK, Murphy-Pace K, Bergstrom N. Systematic review of current executive function measures in adults with and without cognitive impairments. International Journal of Evidence-Based Healthcare. 2010; 8(3):110–125. DOI: 10.1111/j.1744-1609.2010.00170.x [PubMed: 21199379]

Pineda DA, Merchan V. Executive function in young Colombian adults. International Journal of Neuroscience. 2003; 113(3):397–410. DOI: 10.1080/00207450390162164 [PubMed: 12803141]

Pribram KH. The primate frontal cortex – executive of the brain. In: Pribram KH, Luria AR, editorsPsychophysiology of the frontal lobes. New York: Academic Press; 1973. 293–314.

R Core Team. R: A language and environment for statistical computing [computer software]. Vienna, Austria: R Foundation for Statistical Computing; 2013.

Rabin LA, Barr WB, Burton LA. Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. Archives of Clinical Neuropsychology. 2005; 20(1):33–65. DOI: 10.1016/j.acn.2004.02.005 [PubMed: 15620813]

Rabin LA, Paolillo E, Barr WB. Stability in test-usage practices of clinical neuropsychologists in the United States and Canada over a 10-year period: A follow-up survey of INS and NAN members. Archives of Clinical Neuropsychology. 2016; 31:206–230. DOI: 10.1093/arclin/acw007 [PubMed: 26984127]

Reynolds CR, Horton AM. Assessing executive functions: A life- span perspective. Psychology in the Schools. 2008; 45(9):875–892. DOI: 10.1002/pits.20332

Roberts N, Grover V. Theory development in information systems research using structural equation modeling: Evaluation and recommendations. In: Dwivedi YK, editorHandbook of research on contemporary theoretical models in information systems. Hershey, Pennsylvania: Information Science Reference; 2009. 77–94.

Rodríguez-Aranda C, Sundet K. The frontal hypothesis of cognitive aging: Factor structure and age effects on four frontal tests among healthy individuals. The Journal of Genetic Psychology. 2006; 167(3):269–287. DOI: 10.3200/GNTP.167.3.269-287 [PubMed: 17278416]

Roebers CM, Röthlisberger M, Neuenschwander R, Cimeli P, Michel E, Jäger K. The relation between cognitive and motor performance and their relevance for children's transition to school: a latent variable approach. Human Movement Science. 2014; 33:284–297. DOI: 10.1016/j.humov.2013.08.011 [PubMed: 24289983]

Rose SA, Feldman JF, Jankowski JJ. Modeling a cascade of effects: The role of speed and executive functioning in preterm/full-term differences in academic achievement. Developmental Science. 2011; 14(5):1161–1175. DOI: 10.1111/j.1467-7687.2011.01068.x [PubMed: 21884331]

*. Rose SA, Feldman JF, Jankowski JJ. Implications of infant cognition for executive functions at age 11. Psychological Science. 2012; 23(11):1345–1355. DOI: 10.1177/0956797612444902 [PubMed: 23027882]

Rosseel Y. lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software. 2012; 48(2):1–36. http://dx.doi.org/10.18637/jss.v048.i02.

Royall DR, Palmer RF. "Executive functions" cannot be distinguished from general intelligence: Two variations on a single theme within a symphony of latent variance. Frontiers in Behavioral Neuroscience. 2014; 8(369):1–10. DOI: 10.3389/fnbeh.2014.00369 [PubMed: 24478648]

Royall DR, Lauterbach EC, Cummings JL, Reeve A, Rummans TA, Kaufer DI, … Coffey CE. Executive control function: a review of its promise and challenges for clinical research. A report from the Committee on Research of the American Neuropsychiatric Association. The Journal of Neuropsychiatry and Clinical Neurosciences. 2002; 14(4):377–405. DOI: 10.1176/jnp.14.4.377 [PubMed: 12426407]

Salthouse TA, Atkinson TM, Berish DE. Executive functioning as a potential mediator of age-related cognitive decline in normal adults. Journal of Experimental Psychology: General. 2003; 132(4):566–594. DOI: 10.1037/0096-3445.132.4.566 [PubMed: 14640849]

Samyn V, Roeyers H, Bijttebier P, Rosseel Y, Wiersema JR. Assessing effortful control in typical and atypical development: Are questionnaires and neuropsychological measures interchangeable? A latent-variable analysis. Research in Developmental Disabilities. 2015; 36:587–599. DOI: 10.1016/j.ridd.2014.10.018

Sanchez-Cubillo I, Perianez JA, Adrover-Roig D, Rodriguez-Sanchez JM, Rios-Lago M, Tirapu J, Barcelo F. Construct validity of the Trail Making Test: Role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. Journal of the International Neuropsychological Society. 2009; 15(3):438–450. DOI: 10.1017/S1355617709090626 [PubMed: 19402930]

Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. International Journal of Epidemiology. 2007; 36(3):666–676. DOI: 10.1093/ije/dym018 [PubMed: 17470488]

Sapkota S, Vergote D, Westaway D, Jhamandas J, Dixon RA. Synergistic associations of catechol-O-methyltransferase and brain-derived neurotrophic factor with executive function in aging are selective and modified by apolipoprotein E. Neurobiology of Aging. 2015; 36(1):249–256. DOI: 10.1016/j.neurobiolaging.2014.06.020 [PubMed: 25107496]

Scherer R, Tiemann R. Evidence on the effects of task interactivity and grade level on thinking skills involved in complex problem solving. Thinking Skills and Creativity. 2014; 11:48–64. DOI: 10.1016/j.tsc.2013.10.003

Schmidt M. Hit or miss? Insight into executive functions. Journal of the International Neuropsychological Society. 2003; 9(6):962–964. DOI: 10.1017/S1355617703230162

Schoemaker K, Bunte T, Wiebe SA, Espy KA, Deković M, Matthys W. Executive function deficits in preschool children with ADHD and DBD. Journal of Child Psychology and Psychiatry. 2012; 53(2):111–119. DOI: 10.1111/j.1469-7610.2011.02468.x [PubMed: 22022931]

Schreiber JB, Nora A, Stage FK, Barlow EA, King J. Reporting structural equation modeling and confirmatory factor analysis results: A review. The Journal of Educational Research. 2006; 99(6): 323–338. DOI: 10.3200/JOER.99.6.323-338

Scott JC, Matt GE, Wrocklage KM, Crnich C, Jordan J, Southwick SM, … Schweinsburg BC. A quantitative meta-analysis of neurocognitive functioning in posttraumatic stress disorder. Psychological Bulletin. 2015; 141(1):105–140. DOI: 10.1037/a0038039 [PubMed: 25365762]

Senn TE, Espy KA, Kaufmann PM. Using path analysis to understand executive function organization in preschool children. Developmental Neuropsychology. 2004; 26(1):445–464. DOI: 10.1207/s15326942dn2601_5 [PubMed: 15276904]

Shahabi SR, Abad FJ, Colom R. Short-term storage is a stable predictor of fluid intelligence whereas working memory capacity and executive function are not: A comprehensive study with Iranian schoolchildren. Intelligence. 2014; 44:134–141. DOI: 10.1016/j.intell.2014.04.004

Shao Z, Janse E, Visser K, Meyer AS. What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults; Frontiers in Psychology. 2014. 5

Shing YL, Lindenberger U, Diamond A, Li SC, Davidson MC. Memory maintenance and inhibitory control differentiate from early childhood to adolescence. Developmental Neuropsychology. 2010; 35(6):679–697. DOI: 10.1080/87565641.2010.508546 [PubMed: 21038160]

Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science. 2011; 22(11):1359–1366. DOI: 10.1177/0956797611417632 [PubMed: 22006061]

Snyder HR. Major depressive disorder is associated with broad impairments on neuropsychological measures of executive function: A meta-analysis and review. Psychological Bulletin. 2013; 139(1):81–132. DOI: 10.1037/a0028727 [PubMed: 22642228]

Spiess MA, Meier B, Roebers CM. Prospective memory, executive functions, and metacognition are already differentiated in young elementary school children: Evidence from latent factor modeling. Swiss Journal of Psychology. 2015; 74(4):229–241. DOI: 10.1024/1421-0185/a000165

St Clair-Thompson HL, Gathercole SE. Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. The Quarterly Journal of Experimental Psychology. 2006; 59(4):745–759. DOI: 10.1080/17470210500162854 [PubMed: 16707360]

Stuss DT, Levine B. Adult clinical neuropsychology: Lessons from studies of the frontal lobes. Annual Review of Psychology. 2002; 53(1):401–433. DOI: 10.1146/annurev.psych.53.100901.135220

Testa R, Bennett P, Ponsford J. Factor analysis of nineteen executive function tests in a healthy adult population. Archives of Clinical Neuropsychology. 2012; 27:213–224. DOI: 10.1093/arclin/acr112 [PubMed: 22314610]

Teuber HL. Unity and diversity of frontal lobe functions. Acta Neurobiologiae Experimentalis. 1972; 32:615–656. [PubMed: 4627626]

Thibeau S, McFall GP, Wiebe SA, Anstey KJ, Dixon RA. Genetic factors moderate everyday physical activity effects on executive functions in aging: Evidence from the Victoria Longitudinal Study. Neuropsychology. 2016; 30(1):6–17. DOI: 10.1037/neu0000217 [PubMed: 26710092]

Thomas ML, Brown GG, Gur RC, Moore TM, Patt VM, Nock MK, Naifeh JA, Heeringa S, Ursano RJ, Stein MB. on behalf of the Army STARRS Collaborators. Measurement of latent cognitive abilities involved in concept identification learning. Journal of Clinical and Experimental Neuropsychology. 2015; 37(6):653–669. DOI: 10.1080/13803395.2015.1042358 [PubMed: 26147832]

Tucker-Drob EM, Salthouse TA. Methods and measures: Confirmatory factor analysis and multidimensional scaling for construct validation of cognitive abilities. International Journal of Behavioral Development. 2009; 33(3):277–285. DOI: 10.1177/0165025409104489 [PubMed: 20963182]

Tuleya LG, editorThesaurus of psychological index terms. 11. Washington, DC: American Psychological Association; 2009. Rev

Unsworth N. On the division of working memory and long-term memory and their relation to intelligence: A latent variable approach. Acta Psychologica. 2010; 134(1):16–28. DOI: 10.1016/j.actpsy.2009.11.010 [PubMed: 20022311]

Unsworth N, Spillers GJ, Brewer GA. Variation in verbal fluency: A latent variable analysis of clustering, switching, and overall performance. The Quarterly Journal of Experimental Psychology. 2010; 64(3):447–466. DOI: 10.1080/17470218.2010.505292 [PubMed: 20839136]

*. Usai MC, Viterbori P, Traverso L, De Franchis V. Latent structure of executive function in five-and six-year-old children: a longitudinal study. European Journal of Developmental Psychology. 2014; 11(4):447–462. DOI: 10.1080/17405629.2013.840578

*. van der Sluis S, de Jong PF, van der Leij A. Executive functioning in children, and its relations with reasoning, reading, and arithmetic. Intelligence. 2007; 35(5):427–449. DOI: 10.1016/j.intell.2006.09.001

Van der Ven SHG, Kroesbergen EH, Boom J, Leseman PPM. The development of executive functions and early mathematics: A dynamic relationship. British Journal of Educational Psychology. 2012; 82(1):100–119. DOI: 10.1111/j.2044-8279.2011.02035.x [PubMed: 22429060]

*. van der Ven SHG, Kroesbergen EH, Boom J, Leseman PP. The structure of executive functions in children: A closer examination of inhibition, shifting, and updating. British Journal of Developmental Psychology. 2013; 31(1):70–87. DOI: 10.1111/j.2044-835X.2012.02079.x [PubMed: 23331107]

Vandenbroucke JP, Von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, … Egger M. Strengthening the reporting of observational studies in epidemiology (STROBE): Explanation and elaboration. Annals of internal medicine. 2007; 147(8):163–194. DOI: 10.7326/0003-4819-147-8-200710160-00010-w1

Vaughan L, Giovanello K. Executive function in daily life: Age-related influences of executive processes on instrumental activities of daily living. Psychology and Aging. 2010; 25(2):343–355. DOI: 10.1037/a0017729 [PubMed: 20545419]

Visu-Petra L, Benga O, Miclea M. Dimensions of attention and executive functioning in 5-to 12-years-old children: Neuropsychological assessment with the NEPSY battery. Cognition, Brain, Behavior. 2007; 11(3):585–608.

Visu-Petra L, Cheie L, Benga O, Miclea M. The structure of executive functions in preschoolers: An investigation using the NEPSY battery. Procedia-Social and Behavioral Sciences. 2012; 33:627–631. DOI: 10.1016/j.sbspro.2012.01.197

Viterbori P, Usai MC, Traverso L, De Franchis V. How preschool executive functioning predicts several aspects of math achievement in Grades 1 and 3: A longitudinal study. Journal of Experimental Child Psychology. 2015; 140:38–55. DOI: 10.1016/j.jecp.2015.06.014 [PubMed: 26218333]

*. Was CA. Further evidence that not all executive functions are equal. Advances in Cognitive Psychology. 2007; 3(3):399–407. DOI: 10.2478/v10053-008-0004-5

Watkins MW. Structure of the Wechsler Intelligence Scale for Children—Fourth Edition among a national sample of referred students. Psychological Assessment. 2010; 22(4):782–787. DOI: 10.1037/a0020043 [PubMed: 21133545]

Wechsler D. Wechsler adult intelligence scale. 4. San Antonio, TX: Pearson; 2008.

Wechsler D. Wechsler intelligence scale for children. 5. San Antonio, TX: Pearson; 2014.

Weiss LG, Keith TZ, Zhu J, Chen H. WAIS-IV and clinical validation of the four-and five-factor interpretative approaches. Journal of Psychoeducational Assessment. 2013; 31(2):94–113. DOI: 10.1177/0734282913478030

Weiss LG, Keith TZ, Zhu J, Chen H. WISC-IV and clinical validation of the four-and five-factor interpretative approaches. Journal of Psychoeducational Assessment. 2013; 31(2):114–131. DOI: 10.1177/0734282913478032

Welsh JA, Nix RL, Blair C, Bierman KL, Nelson KE. The development of cognitive skills and gains in academic school readiness for children from low-income families. Journal of Educational Psychology. 2010; 102(1):43–53. DOI: 10.1037/a0016738 [PubMed: 20411025]

Welsh MC, Pennington BF. Assessing frontal lobe functioning in children: Views from developmental psychology. Developmental Neuropsychology. 1988; 4(3):199–230. DOI: 10.1080/87565648809540405

Werner H. The concept of development from a comparative and organismic point of view. In: Harris DB, editorThe concept of development: An issue in the study of human behavior. Minneapolis, MN: Jones Press; 1957. 125–148.

*. Wiebe SA, Espy KA, Charak D. Using confirmatory factor analysis to understand executive control in preschool children: I. Latent structure. Developmental Psychology. 2008; 44(2):575–587. DOI: 10.1037/0012-1649.44.2.575 [PubMed: 18331145]

*. Wiebe SA, Sheffield T, Nelson JM, Clark CA, Chevalier N, Espy KA. The structure of executive function in 3-year-olds. Journal of Experimental Child Psychology. 2011; 108(3):436–452. DOI: 10.1016/j.jecp.2010.08.008 [PubMed: 20884004]

Willoughby MT, Blair CB. The Family Life Project Investigators. Measuring executive function in early childhood: A case for formative measurement. Psychological Assessment. 2016; 28(3): 319–330. DOI: 10.1037/pas0000152 [PubMed: 26121388]

Willoughby MT, Blair CB, Wirth RJ, Greenberg M. The Family Life Project Investigators. The measurement of executive function at age 3 years: psychometric properties and criterion validity of a new battery of tasks. Psychological Assessment. 2010; 22(2):306–317. DOI: 10.1037/a0018708 [PubMed: 20528058]

*. Willoughby MT, Blair CB, Wirth RJ, Greenberg M. The Family Life Project Investigators. The measurement of executive function at age 5: Psychometric properties and relationship to academic achievement. Psychological Assessment. 2012a; 24(1):226.doi: 10.1037/a0025361 [PubMed: 21966934]

Willoughby MT, Wirth RJ, Blair CB. The Family Life Project Investigators. Executive function in early childhood: Longitudinal measurement invariance and developmental change. Psychological Assessment. 2012b; 24(2):418–431. DOI: 10.1037/a0025779 [PubMed: 22023561]

Willoughby M, Blair C. Test-retest reliability of a new executive function battery for use in early childhood. Child Neuropsychology. 2011; 17(6):564–579. DOI: 10.1080/09297049.2011.554390 [PubMed: 21714751]

Willoughby M, Holochwost SJ, Blanton ZE, Blair CB. Executive functions: Formative versus reflective measurement. Measurement: Interdisciplinary Research & Perspectives. 2014; 12(3): 69–95. DOI: 10.1080/15366367.2014.929453

Wolf EJ, Harrington KM, Clark SL, Miller MW. Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. Educational and Psychological Measurement. 2013; 73(6):913–934. DOI: 10.1177/0013164413495237

Wolff M, Krönke KM, Venz J, Kräplin A, Bühringer G, Smolka MN, Goschke T. Action versus state orientation moderates the impact of executive functioning on real-life self-control. Journal of Experimental Psychology: General. 2016; 145(12):1635–1653. DOI: 10.1037/xge0000229 [PubMed: 27736135]

Wu KK, Chan SK, Leung PWL, Liu W, Leung FLT, Ng R. Components and developmental differences of executive functioning for school-aged children. Developmental Neuropsychology. 2011; 36(3): 319–337. DOI: 10.1080/87565641.2010.549979 [PubMed: 21462010]

*. Xu F, Han Y, Sabbagh MA, Wang T, Ren X, Li C. Developmental differences in the structure of executive function in middle childhood and adolescence. PloS one. 2013; 8(10):e77770.doi: 10.1371/journal.pone.0077770 [PubMed: 24204957]

Yeh Z. Role of theory of mind and executive function in explaining social intelligence: A structural equation modeling approach. Aging & Mental Health. 2013; 17(5):527–534. DOI: 10.1080/13607863.201.7258235 [PubMed: 23336440]

Yehene E, Meiran N. Is there a general task switching ability? Acta Psychologica. 2007; 126(3):169–195. DOI: 10.1016/j.actpsy.2006.11.007 [PubMed: 17223059]

Ziegler G, Dahnke R, Winkler AD, Gaser C. Partial least squares correlation of multivariate cognitive abilities and local brain structure in children and adolescents. NeuroImage. 2013; 82:284–294. DOI: 10.1016/j.neuroimage.2013.05.088 [PubMed: 23727321]

Public Significance Statement

Previous research has explored whether executive functions are best described as a single self-regulatory ability (i.e., unity) or a diverse set of abilities related to control over thoughts and behaviors (i.e., diversity). This systematic review identified three abilities most frequently evaluated in psychological research (i.e., inhibition, shifting, and updating/working memory), and a re-analysis of previous studies identified greater unity of executive functions during childhood and greater diversity arising from adolescence into adulthood.

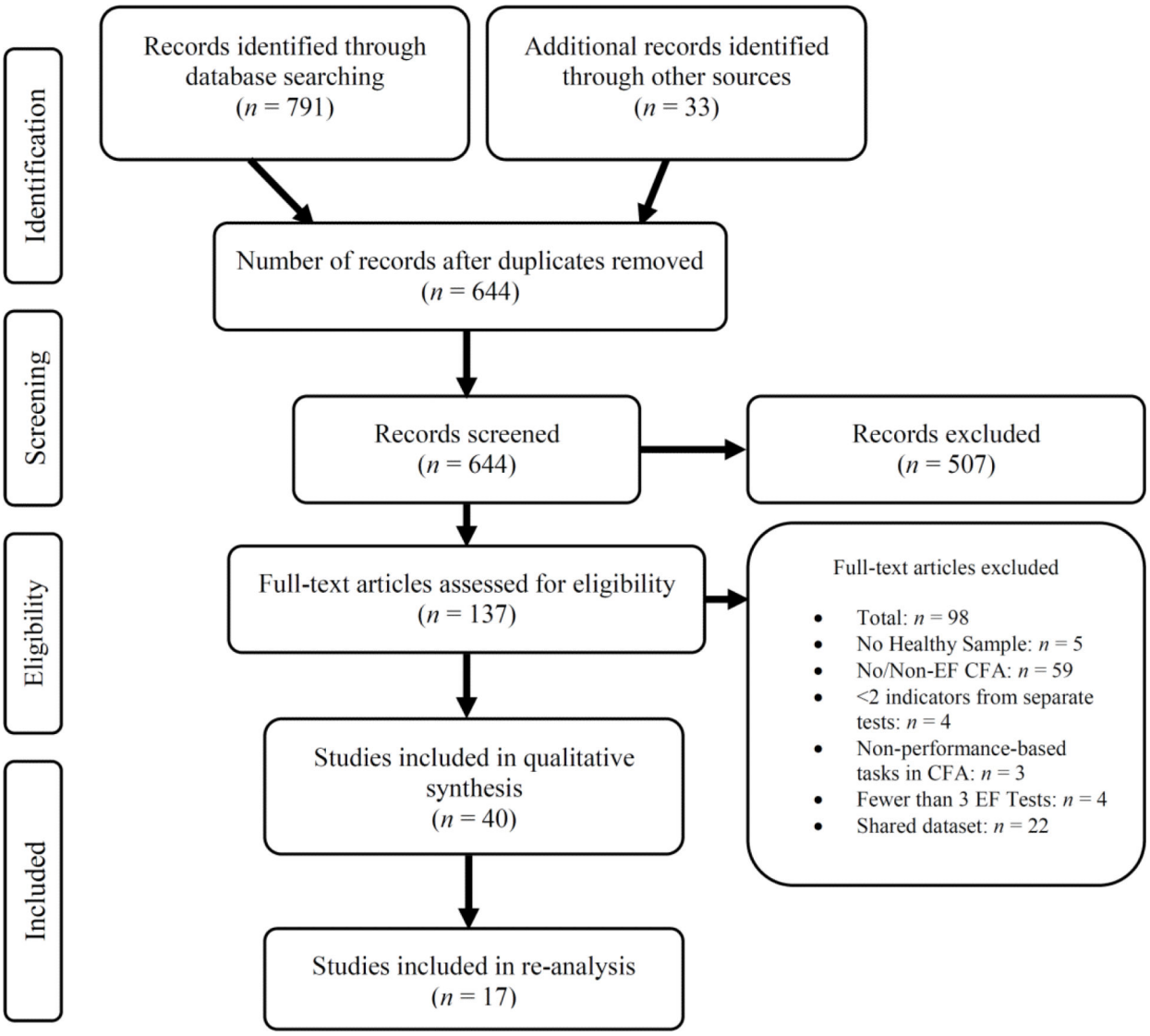

Figure 1.
Flow diagram of systematic review

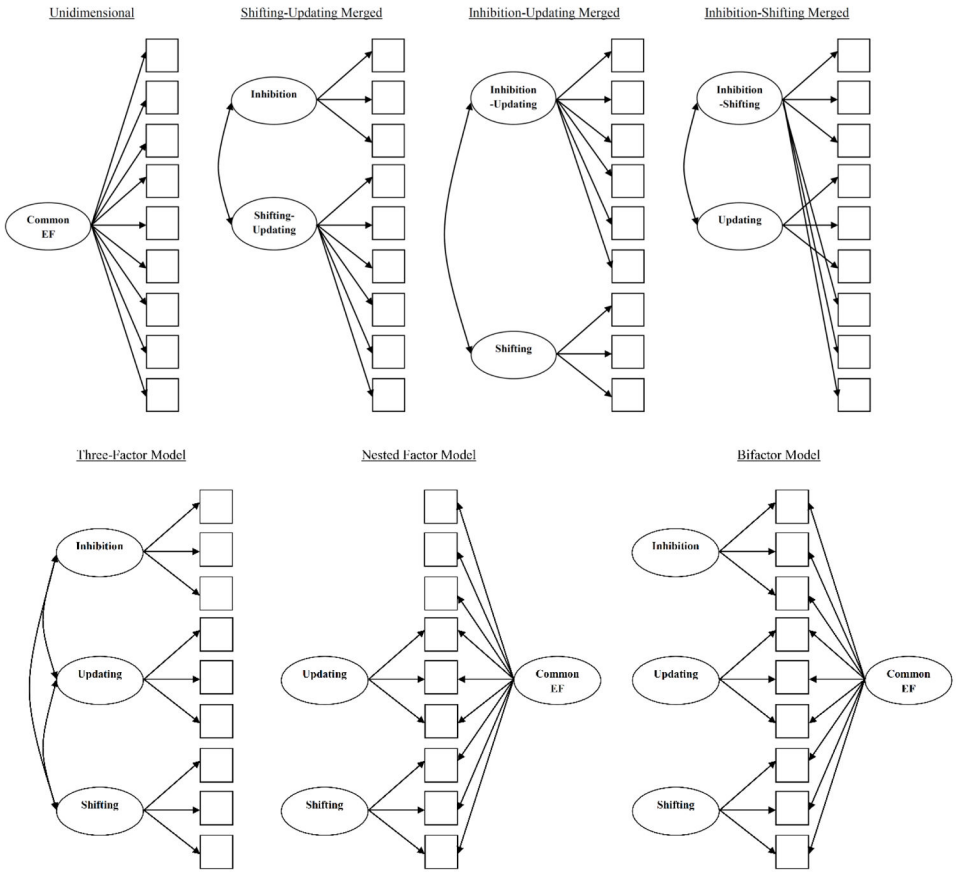

Figure 2.
Diagrams of factor models tested in the re-analysis

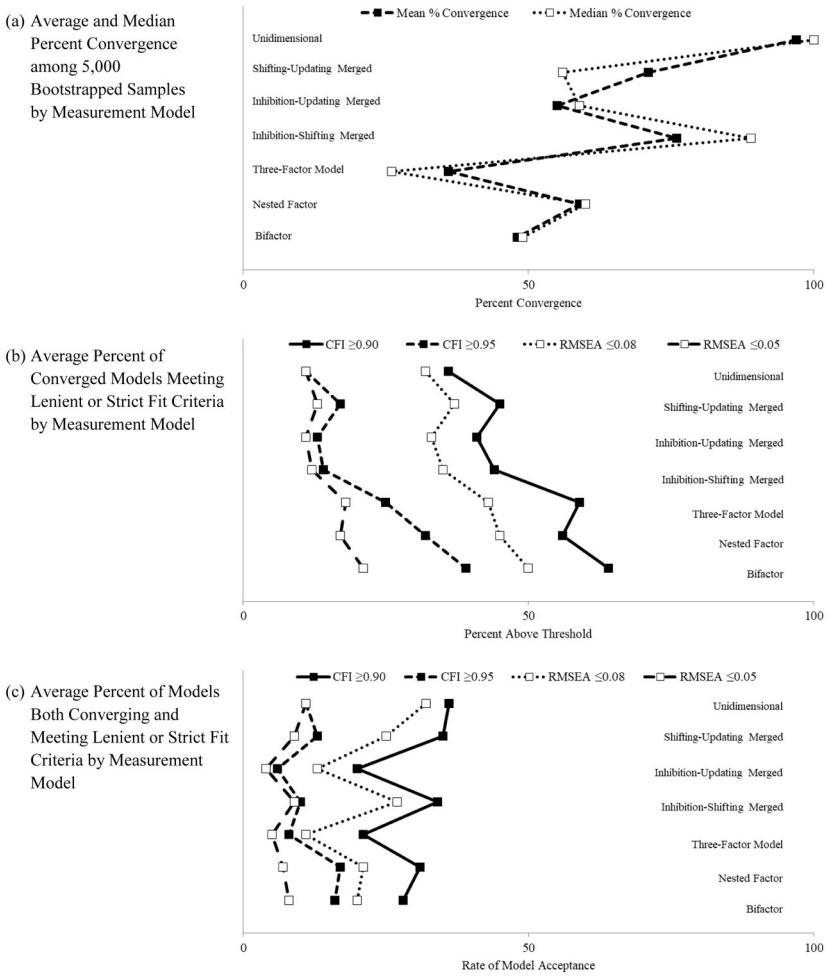

Figure 3.

Child and Adolescent Studies: Forest Plots of Percent Convergence, Percent Meeting Fit Criteria, and Percent Both Converging and Meeting Fit Criteria among 5,000 Bootstrapped Samples

Note. CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; Lenient Fit Criteria = CFI ≥ 0.90 and RMSEA ≤ 0.08; Strict Fit Criteria = CFI ≥ 0.95 and RMSEA ≤ 0.05.

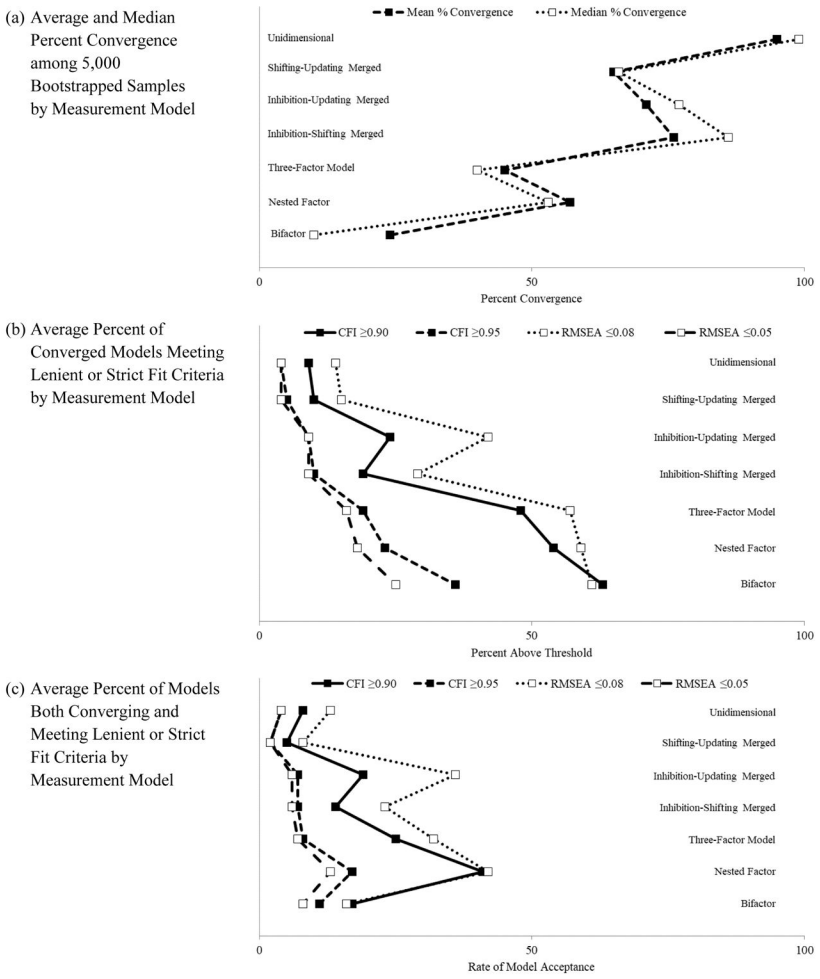

Figure 4.

Adult Studies: Forest Plots of Percent Convergence, Percent Meeting Fit Criteria, and Percent Both Converging and Meeting Fit Criteria among 5,000 Bootstrapped Samples

Note. CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; Lenient Fit Criteria = CFI ≥ 0.90 and RMSEA ≤ 0.08; Strict Fit Criteria = CFI ≥ 0.95 and RMSEA ≤ 0.05.

(a) Percent Model Selection: Average Percent of Bootstrapped Samples where Model Converged and was Selected based on Lenient or Strict ΔCFI or ΔRMSEA Cutoffs

(b) Percent Contingent Model Selection: Among Bootstrapped Samples where the Model Converged, Average Percent of Samples where the Model was Selected based on Lenient or Strict ΔCFI or ΔRMSEA Cutoffs

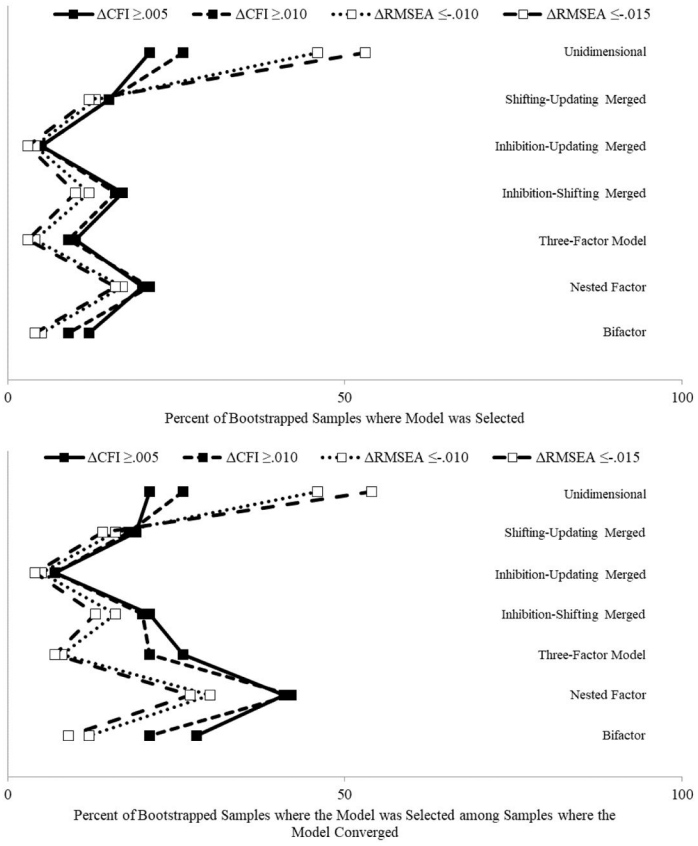

Figure 5.

Child and Adolescent Studies: Forest Plots of the Percent Model Selection and Percent Contingent Model Selection

Note. ΔCFI = Change in Comparative Fit Index; ΔRMSEA = Change in Root Mean Square Error of Approximation; Lenient Change in Fit Cutoffs = ΔCFI ≥ .005 and ΔRMSEA ≤ −.010; Strict Change in Fit Criteria = ΔCFI ≥ .010 and ΔRMSEA ≤ −.015.

(a) Percent Model Selection: Average Percent of Bootstrapped Samples where Model Converged and was Selected based on Lenient or Strict ΔCFI or ΔRMSEA Cutoffs

(b) Percent Contingent Model Selection: Among Bootstrapped Samples where the Model Converged, Average Percent of Samples where the Model was Selected based on Lenient or Strict ΔCFI or ΔRMSEA Cutoffs

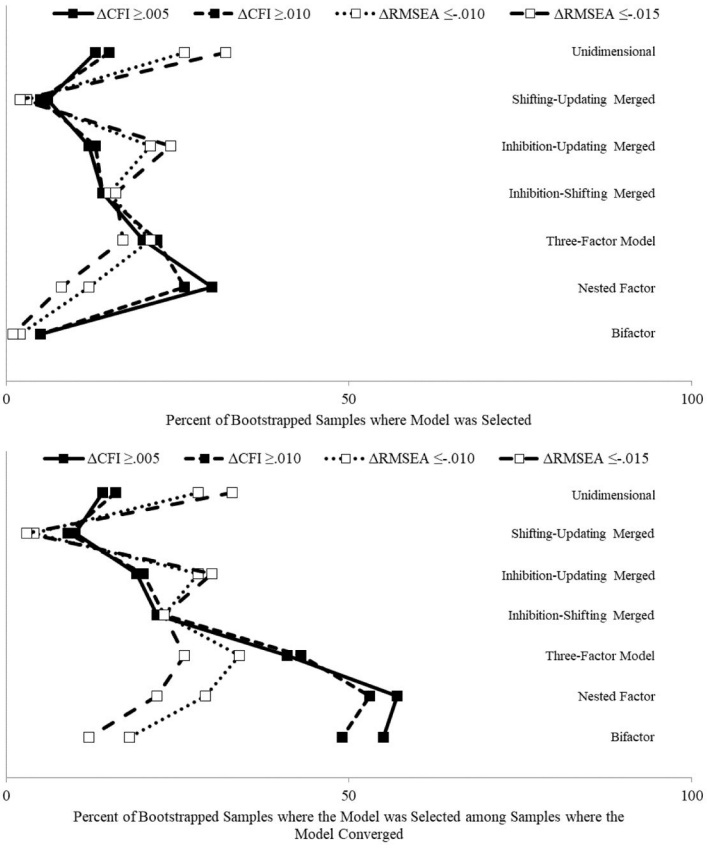

Figure 6.

Adult Studies: Forest Plots of the Percent Model Selection and Percent Contingent Model Selection

Note. ΔCFI = Change in Comparative Fit Index; ΔRMSEA = Change in Root Mean Square Error of Approximation; Lenient Change in Fit Cutoffs = ΔCFI ≥ .005 and ΔRMSEA ≤ −.010; Strict Change in Fit Criteria = ΔCFI ≥ .010 and ΔRMSEA ≤ −.015.

Table 1

Studies Reporting Measurement Models of Executive Functions: Sample Characteristics and Study Quality

Age Group ($\bar{x}$ age)	Author	*N*	Age (years): $\bar{x}$ (*SD*)	Age Range (years)	% Female	% White or Category	Education: $\bar{x}$ (years) or Category	Study Quality*
Preschool (<6 yrs.)	Carlson et al. (2014)	104	4.00 (0.43)	3–5	46.15	80.00	Some preschool	9
	Lerner & Lonigan (2014)	289	4.65 (0.63)	3–5	53.00	57.00	–	8
	Masten et al. (2012)	138	5.77 (0.58)	4.83–6.92	56.50	4.30	–	6
	Miller et al. (2012)+	129	4.17 (0.58)	3–5	39.53	80.00	–	9
	Monette et al. (2015)	272	5.70 (0.34)	–	54.55	–	–	10
	Usai et al. (2014)+	175	5.71 (0.28)	5–6	43.43	–	Kindergarteners	9
	Wiebe et al. (2008)	243	3.92 (1.00)	2–6	55.56	70.37	Preschool	9
	Wiebe et al. (2011)	228	3.01 (0.04)	–	49.56	75.88	–	9
	Willoughby et al. (2012a)	1036	5.03 (0.26)	–	50.00	–	–	8
School-Aged (6–12 yrs.)	Agostino et al. (2010)+	155	10.08 (1.25)	8–13	56.00	–	21.94% Grade 3; 25.81% Grade 4; 28.39% Grade 5; 23.87% Grade 6	7
	Arán-Filippetti (2013)[1]+	LSES: 124 MSES: 124	–	8–12	–	–	Some grade school	8
	Brocki & Tillman (2014)	114	9.32 (2.31)	5–14	47.00	68.00	–	10
	Brydges et al. (2012)+	215	8.33 (1.08)	7–9	48.84	–	–	9
	Duan et al. (2010)+	61	11.88 (0.65)	11–12	44.27	From “Chinese schools”	–	7
	Lambek & Shevlin (2011)[2]	164	–	7–12	47.00	–	–	5
	Lee et al. (2012)	163	6.90 (0.31)	–	49.69	“Mainly ethnic Chinese”	–	8
	Lee et al. (2013)[3]	332	–	All 8	–	–	–	9
	Lehto et al. (2003)+	103	10.50 (1.30)	8–13	44.44	–	–	8
	Rose et al. (2012): Full-Term[4]+	131	11.14 (0.35)	10.5–12.5	49.30	16.10	–	8
	Rose et al. (2012): Pre-Term[4]+	–	11.18 (0.44)	10.4–12.1	–	11.40	–	
	van der Sluis et al. (2007)	172	10.67 (0.72)	9–12	51.16	–	58.14% - Grade 4; 41.86% - Grade 5	10
	van der Ven et al. (2013)	211	6.42 (0.37)	5–7	47.87	–	–	11

Age Group ($\bar{x}$ age)	Author	*N*	Age (years): $\bar{x}$ (*SD*)	Age Range (years)	% Female	% White or Category	Education: $\bar{x}$ (years) or Category	Study Quality*
	Xu et al. (2013)[5+]	140	8.78 (0.57)	7–9	47.86	"Chinese"	–	8
		165	11.59 (0.88)	10–12	49.09	"Chinese"	–	
Adolescents (13–17 yrs.)	Friedman et al. (2011) [+]	813	17.30 (0.60)	15–20	51.66	–	–	11
	Lambek & Shevlin (2011)[2]	75	–	13–16	45.00	–	–	5
	Xu et al. (2013)[5+]	152	14.41 (0.86)	13–15	50.00	"Chinese"	–	8
Adults (18–59 yrs.)	Chuderski et al. (2012)	160	21.90 (2.70)	15–35	61.25	–	–	8
	Fleming et al. (2016)[+]	420	22.50	–	50.00	91.00	–	10
	Fournier-Vicente et al. (2008)	180	23.62 (3.24)	18–31	50.00	–	Undergraduates	11
	Ito et al. (2015)[+]	484	19.75 (2.21)	18–42	49.18	86.16	Undergraduates	11
	Klauer et al. (2010)[6] - Study 1[+]	125	23.10 (5.80)	17–57	42.97	–	–	8
	Klauer et al. (2010)[6] - Study 2[+]	118	23.50 (4.20)	18–42	35.25	–	–	9
	McVay & Kane (2012)	258	–	18–35	–	–	Undergraduates	9
	Miyake et al. (2000)[+]	137	–	–	–	–	Undergraduates	9
	Was (2007)	188	25.70	18–56	70.21	–	Undergraduates	8
Older Adults (>60 yrs.)	Adrover-Roig et al. (2012)	122	62.30 (8.40)	48–91	65.00	"Predominantly Caucasian"	11.30	9
	Bettcher et al. (2016)	202	73.68 (6.60)	63–99	50.50	–	17.67	8
	de Frias et al. (2009): CE group[+]	77	66.05 (7.83)	55–86	67.54[7]	–	15.81	7
	de Frias et al. (2009): CN group[+]	276	68.45 (8.65)	54–88	67.54[7]	–	15.29	
	Frazier et al. (2015)	119	73.00 (6.50)	55–99	54.00	–	17.50	1
	Hedden & Yoon (2006)[+]	121	72.24 (4.28)	63–82	57.02	–	15.69	8
	Hull et al. (2008)[+]	100	60.24 (5.58)	51–74	80.00	–	25%-High School; 2% -Associates; 49% - Bachelor's; 24% - Advanced degree	7
	Vaughan & Giovanello (2010)	95	74.40 (6.40)	60–90	56.00	85.00	16.10	8
Multiple Ages	Huizinga et al. (2006)[8]	71	7.2	6–8	54.93	–	0.56	4

Age Group ($\bar{x}$ age)	Author	*N*	Age (years): $\bar{x}$ (*SD*)	Age Range (years)	% Female	% White or Category	Education: $\bar{x}$ (years) or Category	Study Quality*
		108	11.2	10–12	57.41	–	3.92	
		111	15.3	14–16	52.25	–	7.2	
		94	20.8	18–26	76.6	–	10.55	
	Pettigrew & Martin (2014)[9]	102	21.00 (3.10)	18–32	–	–	14.00	10
		60	71.00 (5.00)	64–87	–	–	16.00	

Note. CE = Cognitively Elite; CN = Cognitively Normal; LSES = Low Socioeconomic Status Group; MSES = Medium Socioeconomic Status Group.

*Study Quality based on items listed under Data Extraction subheading in the Methods section.

+Indicates inclusion in the re-analysis.

[1]Arán-Filippetti (2013) reported confirmatory factor analyses for two separate groups (LSES and MSES).

[2]Lambek & Shevlin (2011) reported two separate confirmatory factor analyses for child and adolescent groups.

[3]Lee et al. (2013) provided a far more comprehensive span of ages; however, due to its sequential cohort-design, there was significant overlap between participants at different ages. In order to ensure that the same individuals were not represented twice in the systematic review, and to increase comparability with other designs, only the cross-section with the greatest amount of participants is considered in the current review and presented in the current table. There is a significant amount of demographic information provided in the original article for the cohorts at baseline; however, the data was not available for the cohort selected for consideration in the current review.

[4]Demographic statistics for Rose et al. (2012) reported separately for full-term and pre-term participants, but only one confirmatory factor analysis was run using the full sample. Some statistics were pulled from Rose et al. (2011), which used the same participant sample.

[5]Xu et al. (2013) reported three separate confirmatory factor analyses for two child and one adolescent group.

[6]Klauer et al. (2010) reported two separate studies, involving separate samples and separate confirmatory factor analyses.

[7]de Frias et al. (2009) did not report separate gender breakdowns for their CE and CN subgroups, so the value reported above was from full sample.

[8]Huizinga et al. (2006) reported demographics for four separate age groups, and reported fit indices corresponding to a configural invariance model across age groups.

[9]Pettigrew & Martin (2014) merged their young and old participants into one group for their confirmatory factor analysis, but did not report separate demographic characteristics for the merged group.

Table 2

Studies Reporting Measurement Models of Executive Functions: Fit Indices and Latent Constructs

Age Group	Author	χ^2 (*p*)	*df*	CFI	RMSEA	$\hat{\pi}^1$	Accepted Model	EF-related Factors	Factors tested, but removed/merged	Non-EF Factors in Model
Preschool (<6 yrs.)	Carlson et al. (2014)[2]	24.25 (0.15); 24.56 (0.14)	18; 18	0.96; 0.96	0.06; 0.06	0.15	Both One-and Two-Factor Acceptable	EF; Conflict EF, Delay EF	Conflict and Delay EF merged to form EF	None
	Lerner & Lonigan (2014)	55.60 *	33	0.97	0.05	0.78	Two-Factor	Inhibitory Control, WM	Inhibitory Control - Suppression and Inhibitory Control - Conflict merged to form Inhibitory Control	None
	Masten et al. (2012)	–	–	0.97	0.04	–	One-Factor	EF	Hot EF and Cool EF merged to form EF	IQ
	Miller et al. (2012) +	43.41	42	1.00	0.02	0.39	Two-Factor	Inhibition, WM	WM and Set-Shifting merged to form WM	None
	Monette et al. (2015)	60.92 (0.59)	64	1.00	0.00	0.88	Two-Factor	Inhibition, Flexibility-WM	Flexibility and WM merged to form Flexibility-WM	Speed (Control Factor)
	Usai et al. (2014) +	9.48	8	0.98	0.03	0.20	Two-Factor	Inhibition, WM-Shifting	WM and Shifting merged to form WM-Shifting	None
	Wiebe et al. (2008)	31.14 (0.27)	27	0.99	0.03	0.52	One-Factor	EF	WM, Interference from Distractors, and Proactive Interference merged to form EF	None
	Wiebe et al. (2011)	14.84 (0.39)	14	0.99	0.02	0.36	One-Factor	EF	Inhibition and WM and merged to form EF	None
	Willoughby et al. (2012a)	6.30 (0.71)	9	1.00	0.00	0.96	One-Factor	EF	Inhibitory Control/Attention Shifting and WM merged to form EF	None
School-Aged (6–12 yrs.)	Agostino et al. (2010) +	33.26 (0.23)	28	0.98	0.035	0.32	Three-Factor	Inhibition, Updating, Shifting	None	Mental-Attentional Capacity
	Arán-Filippetti (2013)[3]: LSES +	30.65 (0.13)	23	0.97	0.05	0.18	Three-Factor	Inhibition, WM, Cognitive Flexibility	None	None
	Arán-Filippetti (2013)[3]: MSES +	21.35 (0.56)	23	1.00	0.00	0.18	Three-Factor	Inhibition, WM, Cognitive Flexibility	None	None
	Brocki & Tillman (2014)	16.78 (0.61)	19	1.00	<.001	0.15	Two-Factor	Inhibition, WM	None	None
	Brydges et al. (2012) +	20.11 (0.45)	20	1.00	0.01	0.33	One-Factor	EF	Inhibition, Shifting and WM merged to form EF	None
	Duan et al. (2010)[4]+	8.04 (0.24)	–	0.98	0.08	–	Three-Factor	Inhibition, Updating, Shifting	None	None
	Lambek & Shevlin (2011)[5]	3.07 (0.80)	6	1.00	0.00	0.10	Three-Factor	Inhibition, Verbal WM, Visuospatial WM	None	None
	Lee et al. (2012)	135.14 *	86	0.94	0.06	0.66	Two-factor	Inhibition/Switch, Updating	Inhibition and Switching merged to form Inhibition/Switch	Reaction Time, Flanker Task, Simon Task (All Control Factors)
	Lee et al. (2013)[6]	145.47 *	68	0.97	0.06	0.99	Two-Factor	Inhibition/Switch, Updating	Inhibition and Switching merged to form Inhibition/Switch	Control conditions for each task predicted indicators from the same task
	Lehto et al. (2003) +	13.73	16	1.00	–	0.13	Three-Factor	Inhibition, WM, Shifting	None	None

Age Group	Author	χ^2 (*p*)	*df*	CFI	RMSEA	$\hat{\pi}^1$	Accepted Model	EF-related Factors	Factors tested, but removed/merged	Non-EF Factors in Model
	Rose et al. (2012)[+]	41.88 (0.11)	32	0.96	0.05	0.32	Three-Factor	Inhibition, WM, Shifting	WM: Storage and WM: Storage and Processing merged to form WM	None
	van der Sluis et al. (2007)	190.99[*]	122	0.95	0.05	0.78	Nested Factor Model	Updating, Shifting	Inhibition merged with Naming	Naming (Control factor)
	van der Ven et al. (2013)	173.43[*] (0.00)	121	0.96	0.05	0.93	Two-Factor	Inhibition/Shifting, Updating	Inhibition and Shifting merged to form Inhibition/Shifting	Verbal Speed, Motor Speed (Both Control Factors)
	Xu et al. (2013)[7]: Ages 7–9[+]	15.65 (0.34)	14	0.95	0.03	0.19	One-Factor	EF	Inhibition, Updating WM, and Shifting merged to form EF	None
	Xu et al. (2013)[7]: Ages 10–12[+]	19.24 (0.30)	14	0.95	0.05	0.19	One Factor	EF	Inhibition, Updating WM, and Shifting merged to form EF	None
Adolescents (13–17 yrs.)	Friedman et al. (2011)[+]	53.56[*]	21	0.96	0.04	0.99	Nested Factor Model	EF, Updating, Shifting	Inhibition merged with EF	None
	Lambek & Shevlin (2011)[5]	3.99 (0.67)	6	1.00	0.00	0.08	Three-Factor	Inhibition, Verbal WM, Visuospatial WM	None	None
	Xu et al. (2013)[7]: Ages 13 to 15[+]	15.72 (0.15)	11	0.95	0.05	0.15	Three-Factor	Inhibition, Updating WM, Shifting	None	None
Adults (18–59 yrs.)	Chuderski et al. (2012)	70.20	60	0.96	0.02	0.52	Four-Factor	Attention Control, Interference Resolution, Response Inhibition, Storage Capacity	Updating merged with Storage Capacity	None
	Fleming et al. (2016)[8+]	30.36 (0.09)	21	0.97	0.03	0.74	Nested Factor Model	EF, Updating, Shifting	Inhibition not tested, but indicators included on EF bifactor	None
	Fournier-Vicente et al. (2008)	91.62 (0.17)	80	0.99	0.03	0.83	Five-Factor	Verbal SPC, Visuospatial SPC, Selective Attention, Shifting, Strategic Retrieval	Dual-Task Coordination removed	None
	Ito et al. (2015)[+]	32.01 (0.04)[*]	20	0.98	0.04	0.87	Nested Factor Model	EF, Updating, Shifting	Inhibition not tested, but indicators included on EF bifactor	None
	Klauer et al. (2010)[9] - Study 1[+]	12.82 (0.38)	12	0.98	0.02	0.15	Two-Factor	Inhibition/WM, Switching	Inhibition merged with WM to form Inhibition/WM	None
	Klauer et al. (2010)[9] - Study 2[+]	41.09 (0.13)	32	0.94	0.05	0.20	Three-Factor	Inhibition, WM, Switching	None	None
	McVay & Kane (2012)	194.51[*]	126	0.92	0.05	0.99	Two-Factor	WM Capacity, Attention Control	None	Task-unrelated thoughts, Reading Comprehension
	Miyake et al. (2000) [+]	20.29 (0.65)	24	1.00	–	0.28	Three-Factor	Inhibition, Updating, Shifting	None	None
	Was (2007)	12.61 (0.13)	8	0.97	0.06	0.20	Two-Factor	Inhibiting, Updating	None	None
Older Adults (>60 yrs.)	Adrover-Roig et al. (2012)	19.86 (0.70)	24	1.00	0.00	0.18	Three-Factor	WM, Shifting, Access	Inhibition and Updating merged to form WM	None
	Bettcher et al. (2016)	144.12[*]	95	0.96	0.05	0.88	Two-Factor	Shifting/Inhibition, Updating/WM	Mental Set-shifting and Inhibition merged to form Shifting/Inhibition	Speed (Control Factor)
	de Frias et al. (2009)[10]: CE Group[+]	6.53 (0.69); 5.55 (0.48)	9; 6	1.00; 0.98	0.00; 0.00	0.11	Three-Factor	Inhibition, Updating, Shifting	None	None
	de Frias et al. (2009)[10]: CN Group[+]	17.11 (0.05); 5.11 (0.53)	9; 6	0.94; 1.00	0.06; 0.00	0.32	Two-factor	Not specified for untested two-factor model	None	None

Age Group	Author	χ^2 (*p*)	*df*	CFI	RMSEA	$\hat{\pi}$[1]	Accepted Model	EF-related Factors	Factors tested, but removed/merged	Non-EF Factors in Model
Multiple Ages	Frazier et al. (2015)	–	–	–	–	–	Two-Factor	WM, Cognitive Control	Inhibition merged with Set-Shifting to form Cognitive Control	Processing Speed (Control)
	Hedden & Yoon (2006)[+]	115.09	125	1.00	0.00	0.50	Two-Factor	Shifting/Updating, Resistance to Proactive Interference	Shifting and Updating merged to form Shifting/Updating; Prepotent Response Inhibition merged with Speed	Verbal Memory, Visual Memory, Speed
	Hull et al. (2008)[+]	17.76	14	–	0.05	0.13	Two-Factor	Updating, Shifting	Inhibition removed	None
	Vaughan & Giovanello (2010)	30.23 (0.18)	24	0.97	0.05	0.18	Three-factor	Inhibition, Updating, Task Switching	None	None
	Huizinga et al. (2006)[11]	139.34[*]	67	–	–	–	Two-Factor	Stop-Signal Inhibition, Eriksen Flanker Inhibition, Stroop Inhibition, WM, Shifting	Inhibition split into three single-item factors	Basic Speed (Control)
	Pettigrew & Martin (2014)[12]	55.23	42	0.95	0.04	.39	Two-Factor	WM, Interference Resolution	Response-distractor Inhibition and Resistance to Proactive Interference Merged to form Interference Resolution	Age (Control)

Note. CE = Cognitively Elite; CFI = Comparative Fit Index; CN = Cognitively Normal; EF = Executive Function; LSES = Low Socioeconomic Status Group; MSES = Medium Socioeconomic Status Group; RMSEA = Root Mean Square Error of Approximation; SPC = Storage and Processing Coordination; WM = Working Memory; $\hat{\pi}$ = Estimated Power.

[*] Indicates a significant χ^2 test of model fit (p<.05).

[+] Indicates inclusion in the re-analysis.

[1] Power was estimated based on tables provided by Hancock (2006) for post-hoc power analyses of model fit. The values provided herein were based on tables for models with an RMSEA=.00. Because Hancock provided *n* or *df* values in increments of 50 and 5, respectively, the *n* and *df* values from the studies included in the systematic review were rounded to the nearest increments. For models that did not report their *df*, a power value was not estimated, and the studies reporting these models were provided no points for the power criterion of the study quality scale.

[2] Carlson et al. (2014) did not report a preference for either their one- or two-factor model, and the results for both models are reported here, with the one-factor fit indices coming before the semicolon and the two-factor fit indices coming after the semicolon.

[3] Arán-Filippetti (2013) reported confirmatory factor analyses for two separate groups (LSES and MSES).

[4] Duan et al. (2010) reported a χ^2/df value, which is reported here in place of a χ^2 value.

[5] Lambek & Shevlin (2011) reported two separate confirmatory factor analyses for child and adolescent groups.

[6] Lee et al. (2013) used a sequential cohort design, where they recruited participant at different baseline ages and assessed them longitudinally over the course of four years. Consequently, the summary data and fit indices provided for each age group involved a variable amount of overlap (e.g., children starting at age 8 were combined with children that started at age 5 that had already completed three past annual waves of data collection). In turn, only the fit indices from the time point with the largest amount of participants was considered to avoid representing the same individuals twice in the analyses.

[7] Xu et al. (2013) reported three separate confirmatory factor analyses for two child and one adolescent group.

[8] Fleming et al. (2016) and Ito et al. (2015) both included indicators for an inhibition factor, but had these indicators load directly on a general EF bifactor. Guided by previous research, these authors never tested a model including a specific inhibition factor; but because the model included these indicators, inhibition is listed as a factor tested, but removed/merged.

[9] Klauer et al. (2010) reported two separate studies, involving separate samples and separate confirmatory factor analyses.

[10] de Frias et al. (2009) concluded that the data supported a three-factor solution for the CE group (based partially on longitudinal invariance testing); however, the one-factor model fit the data better at Wave 1, and this model was also more parsimonious. The authors concluded a two-factor model best fit the data CN group, although such a model was not tested by the researchers. The fit indices reported herein derive from the one-factor and three-factor models at Wave 1 in their longitudinal design, with the one-factor fit indices coming before the semicolon and the three-factor fit indices coming after the semicolon.

[11] Huizinga et al. (2006) reported fit indices corresponding to a configural invariance model across age groups. Because the sample size was dispersed across the age groups, an estimate of power was not provided for this study.

[12] Pettigrew & Martin (2014) merged their young and old participants into one group for their confirmatory factor analysis.

Table 3

Counts and Frequencies of Constructs represented in Accepted Measurement Models

	Age Group (Age range)	*k*	EF	Inhibition	UWM	Shifting	Inhibition/Shifting	Inhibition/UWM	Shifting/UWM	SR/Access
Counts (*k*)	All Ages	46	11	24	33	20	5	1	3	2
	Preschool (<6 yrs.)	9	5	4	2	0	0	0	2	0
	School-Aged (6–12 yrs.)	15	3	8	12	7	3	0	0	0
	Adolescents (13–17 yrs.)	3	1*	2	3	2	0	0	0	0
	Adults (18–59 yrs.)	9	2*	6	8	6	0	1	0	1
	Older Adults (>60 yrs.)	8	0	3	6	4	2	0	1	1
Frequencies (%)	All Ages	46	23.91	52.17	71.74	43.48	10.87	2.17	6.52	4.35
	Preschool (<6 yrs.)	9	55.56	44.44	22.22	0	0	0	22.22	0
	School-Aged (6–12 yrs.)	15	20.00	53.33	80.00	46.67	20.00	0	0	0
	Adolescents (13–17 yrs.)	3	33.33	66.67	100	66.67	0	0	0	0
	Adults (18–59 yrs.)	9	22.22	66.67	88.89	66.67	0	11.11	0	11.11
	Older Adults (>60 yrs.)	8	0	37.50	75.00	50.00	25.00	0	12.50	12.50

Note.

*The EF factor observed for adolescent and adult samples were general bifactors in models that also included updating and shifting in the same model. EF = Executive Function; SR = Strategic Retrieval; UWM = Updating/Working Memory. The names attributed to similar constructs differed across studies. Selective Attention, Attention Control, Interference Resolution and Response Inhibition, Resistance to Proactive Interference, Inhibitory Control, Inhibiting, and Interference Resolution were subsumed under Inhibition. Updating, WM, WM Capacity, and Storage Capacity were subsumed under Updating/Working Memory. Cognitive Flexibility, Flexibility, Task Switching, and Switching were subsumed by Shifting. Cognitive Control was subsumed under Inhibition/Shifting. Strategic Retrieval and Access were subsumed under Strategic Retrieval/Access. Based on semantic overlap (Packwood et al., 2011), Selective Attention (Fournier-Vicente et al., 2008) and Attention Control (Chuderski et al., 2012) could be subsumed under Shifting; however, the indicators for these factors from both studies were more closely related to Inhibition (e.g., Stroop, Antisaccade), and were thus subsumed under that construct.

Some studies found multiple factors interpretable as sub-dimensions of a common EF-related construct. In these cases, these multiple factors were tallied as representative of a single factor. Specifically, Lambek and Shevlin (2011) found separable Verbal and Visuospatial WM factors, which were tallied as one observation of an Updating/WM factor for each of these authors' two reported samples. Chuderski et al. (2012) found separable Attention Control, Interference Resolution and Response Inhibition factors, which were tallied as one observation of an inhibition factor based on their similarities to this construct based on the authors' conceptual and operational definitions. Lastly, Fournier-Vicente et al. (2008) found separable Verbal and Visuospatial Storage and Processing Coordination, which were tallied as one observation of Updating/WM.

Carlson et al. (2014) did not report a preference for either their one-factor or two-factor model. Based on fit indices, the one-factor was more parsimonious and showed nearly identical fit to the two-factor model. In turn, an EF factor was added to the tally for this study. For the Cognitively Normal group described by de Frias et al. (2009), the authors reported an untested two-factor model as their accepted model. Because this model was untested, it is not clear which factors were represented in this two-factor model, and the results of this group are not represented within this table.