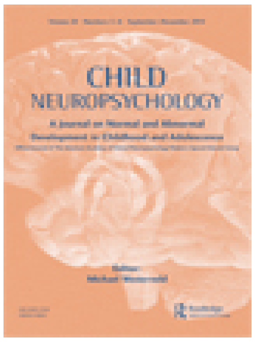

Routledge
Taylor & Francis Group

Child Neuropsychology

A Journal on Normal and Abnormal Development in Childhood and Adolescence

ISSN: 0929-7049 (Print) 1744-4136 (Online) Journal homepage: http://www.tandfonline.com/loi/ncny20

The Childhood Executive Functioning Inventory (CHEXI): Factor structure, measurement invariance, and correlates in US preschoolers

Marie Camerota, Michael T. Willoughby, Laura J. Kuhn & Clancy B. Blair

To cite this article: Marie Camerota, Michael T. Willoughby, Laura J. Kuhn & Clancy B. Blair (2016): The Childhood Executive Functioning Inventory (CHEXI): Factor structure, measurement invariance, and correlates in US preschoolers, Child Neuropsychology, DOI: 10.1080/09297049.2016.1247795

To link to this article: http://dx.doi.org/10.1080/09297049.2016.1247795

Published online: 13 Nov 2016.

CHILD NEUROPSYCHOLOGY, 2016
http://dx.doi.org/10.1080/09297049.2016.1247795

The Childhood Executive Functioning Inventory (CHEXI): Factor structure, measurement invariance, and correlates in US preschoolers

Marie Camerota[a], Michael T. Willoughby[b], Laura J. Kuhn[c] and Clancy B. Blair[d]

[a]Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [b]Education and Workforce Development, RTI International, Research Triangle Park, NC, USA; [c]FPG Child Development Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [d]Department of Applied Psychology, New York University, New York, NY, USA

ABSTRACT

This study tests the factor structure, measurement invariance, and correlates of the Childhood Executive Functioning Inventory (CHEXI) with a large and diverse sample of 3- to 5-year-olds (n = 844). Consistent with previous studies, a two-factor model that distinguishes working memory from inhibition provides the best fit to the observed data. This two-factor model has been shown to demonstrate strong measurement invariance for different subgroups of children (boys vs. girls, high vs. low income). Whereas boys tend to have greater working memory and inhibition difficulties (Cohen's d = 0.15 and 0.20, respectively), children from low-income households tend to have more working memory problems than their peers from high-income households (Cohen's d = 0.25). Finally, correlations between CHEXI scores, examiner reports of child behavior, and child performance on a battery of executive function (EF) tasks were investigated. CHEXI scores were found to be more consistently related to examiner reports of child behavior than child performance on EF tasks. Tthe strengths and weaknesses of the CHEXI as a questionnaire measure of EF are discussed, and directions for future research are suggested.

ARTICLE HISTORY
Received 4 May 2016
Accepted 9 October 2016
Published online
10 November 2016

KEYWORDS
Executive function; Rating scales; Confirmatory factor analysis; Measurement invariance; Performance-based measures

Executive function (EF) is a term which refers to a set of domain-general cognitive processes that support problem-solving, self-management, and goal-directed behavior. Although EF abilities improve across the first two decades of life, early childhood has been identified as an initial period of rapid developmental change (Garon, Bryson, & Smith, 2008). As such, considerable effort has been focused on the measurement of EF in preschool-aged children. While numerous performance-based measures of EF have been developed for use in early childhood (for a review, see Carlson, 2005), far less energy has been devoted to questionnaire measures, despite the fact that clinicians and researchers alike have expressed interest in such measures (Isquith, Roth, & Gioia, 2013).

CONTACT Marie Camerota ✉ marie.camerota@unc.edu Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Campus Box 3270, Chapel Hill, NC 27599-3270, USA

The Behavior Rating Inventory of Executive Function (BRIEF; Gioia, Isquith, Guy, & Kenworthy, 2000) is the most widely used and certainly the most commonly studied questionnaire measure of EF for children and adolescents. The BRIEF includes parent, teacher, and student report forms, and the parent and teacher forms have been modified for use with preschool-aged children in the BRIEF-P (Gioia, Espy, & Isquith, 2003). The BRIEF-P measures observable behaviors in the home and preschool contexts that are presumed to reflect the behavioral manifestations of children's underlying cognitive abilities. One of the key advantages of the BRIEF-P is the availability of normative data, which permits making inferences about whether or not a child's observed behaviors are clinically elevated. As a result, the BRIEF-P is often used in both clinical and research contexts (Isquith et al., 2013). Despite these strengths, the BRIEF-P is relatively long (68 items), takes 10 to 15 minutes to complete, and costs over US$2 per administration, including protocol and associated scoring costs. These factors likely prohibit the use of the BRIEF-P in some low-resource settings.

The Childhood Inventory of Executive Function (CHEXI; Thorell & Nyberg, 2008) is a relatively new questionnaire that represents a potential alternative to BRIEF-P. The CHEXI only has 26 items and so is shorter than the BRIEF-P; it also is available in the public domain (i.e., freely distributed) and has been translated into multiple languages (see http://ww.chexi.se). Although the instrument was designed with four factors in mind (working memory, planning, inhibition, and regulation), the scale's authors found that two latent factors—working memory and inhibition—provide the most parsimonious fit to both parent- and teacher-rated data (Thorell & Nyberg, 2008). These two factors demonstrate acceptable levels of internal consistency (Cronbach's $\alpha > .85$), test–retest reliability ($r > .74$) and criterion validity with attention deficit hyperactivity disorder (ADHD) symptoms of hyperactivity/impulsivity ($r = .27$–$.36$) and inattention ($r = .13$–$.27$; Catale, Meulemans, & Thorell, 2013; Thorell & Nyberg, 2008). Moreover, the two-factor structure has been replicated in other studies (Catale, Lejeune, Merbah, & Meulemans, 2013; Catale, Meulemans, & Thorell, 2013). Despite this consistency in findings, it is worth noting that no study to date has explicitly tested whether the two-factor model provides a statistically significant improvement in fit over the four-factor model using confirmatory factor analysis. Additionally, the studies described above have relied exclusively on samples of European children between the ages of 5 and 11 years. Thus, the first goal of the current investigation was to perform an independent test of the CHEXI factor structure in a sample of preschoolers from the United States (US).

Another question is whether the two-factor solution provides an equally good fit for subgroups. For example, gender (female > male) and income (high income > low income) differences have been observed in previous work involving performance-based measures of EF (Willoughby & Blair, 2015). To the extent that researchers are interested in using questionnaire measures such as the CHEXI to make inferences about group differences in EF (e.g., Thorell, Eninger, Brocki, & Bohlin, 2010), it is important to have some assurance that observed differences in scores are not due to differing measurement properties across groups. Put another way, group differences in mean scores are only interpretable if the measurement properties of an instrument are invariant across groups. As no previous study has tested the measurement invariance of the CHEXI, the second goal of this study is to test whether the CHEXI exhibits measurement invariance across subgroups of children (i.e., boys vs. girls, high income vs. low income). If measurement invariance is established, a corollary question would then be whether group differences are evident by gender or income group.

Despite the widespread interest in questionnaire-based assessments of EF, a recurring finding has been that parent and teacher ratings of behaviors that are presumed to represent EF abilities are only weakly associated with performance-based measures (i.e., median r = .19 across 20 studies; see Toplak, West, & Stanovich, 2013). Two previous studies have reported modest correlations (rs = .20–.30) between parent and teacher ratings on the CHEXI and children's scores on performance-based measures (Catale, Lejeune, et al., 2013; Thorell & Nyberg, 2008). This work suggests that questionnaire- and performance-based assessments are capturing EF abilities at different levels of analysis and are not interchangeable. In order to better understand the relation between questionnaire- and performance-based measures of EF, the current study includes examiner ratings of child behavior during EF task administration. To the extent that child behavior during testing represents another behavioral manifestation of EF ability, and contributes directly to EF task performance, it was expected that examiner ratings would relate to both parents' ratings of children's behavior (per the CHEXI) and children's performance on an EF battery.

In sum, this study provides the first test of the factor structure and measurement invariance of the CHEXI in a US-based sample of preschool-aged children. This study also explores the relation between children's EF-related behaviors (assessed by parents and examiners) and their performance on EF tasks. A better understanding of the magnitude of these relations promises to inform researchers' choices in how to measure EF performance and enhance interpretations of both questionnaire- and performance-based measures of EF.

Method

Participants and Procedure

This participants consist of 846 children who attended preschools in New York or North Carolina. A quota-based sampling procedure was used to guide participant recruitment (Lohr, 2009). Specifically, using data from the 2012 US Census for 0- to 5-year-olds, a distribution of children was established with respect to household income (i.e., poor ≤ 100% US federal poverty threshold for a given household size, near poor > 100% and < 200%, not poor ≥ 200%), race (i.e., Caucasian, African American, Asian, Native American, Pacific Islander), and ethnicity (i.e., Hispanic, non-Hispanic). This information was then used to generate 180 mutually-exclusive cells (3 income levels × 5 races × 2 ethnicities × 3 ages × 2 genders) that served to guide student enrollment. Consistent with a quota sampling approach, the intent was not to recruit exact numbers of children into any given cell (in fact, all eligible and interested children were enrolled); rather, the intent was to recruit a large convenience sample of children that is diverse with respect to race, ethnicity, household income level, age, and gender using the expected cell counts as targets. In general, the target number of children for most race × ethnicity × income level cells was met, with the exception that Caucasian children at all income levels were under-recruited with respect to their target cell counts. The final sample of 846 children varies with respect to age (36% 3-year-olds, 45% 4-year-olds, 19% 5-year-olds), gender (50% male), race (60% Caucasian, 31% African American, 7% Asian, 1% Native American, 1% Pacific Islander), ethnicity (20% Hispanic, 80% non-Hispanic), and household income level (25% poor, 21% near poor, 54% not poor).

Recruitment was undertaken by approaching the directors of center-based preschools that served children in the target age range. Center directors distributed consent forms to parents of children who were in the target range. Interested parents who returned their consent forms to preschools were contacted by research staff for a screening phone call. During the call, parents provided demographic information about themselves and their child, as well as completing the CHEXI items. Children who were outside of the target range (i.e., 3.0–5.9 years of age), who had physical or mental disabilities that prohibited their ability to participate in direct assessments, or who did not speak English were not eligible to participate.

Following the recruitment phone call, children participated in a one-time assessment of EF abilities at their preschool. Of the 924 families who completed recruitment phone calls and were eligible to participate, 92% (n = 846) were tested. Due to the large number of planned assessments (including the *EF Touch* battery), it was infeasible to assign each child to complete all tasks because of anticipated fatigue effects. Therefore, a planned missing design was utilized wherein children were randomized to complete a subset of EF tasks. Moreover, the order in which the assigned tasks were administered was counterbalanced across children. Individual testing sessions were completed by a single research assistant and lasted approximately 30 to 45 minutes for 3-year-olds and 45 to 60 minutes for 4- and 5-year-olds. The preschool centers received US$5 for each completed consent form. Parents received US$40 for the completion of the screening phone call. Children received a small gift for participation. Of the 846 children tested in their preschools, 844 had CHEXI data and were included in the analyses.

Measures

Demographics

During the recruitment phone call, parents provided demographic information about their child and household, including child gender and race, and household income. Although three household income groups were formed for the purposes of recruitment (see above), household income was dichotomized into high income (≥200% of the poverty threshold) and low income (<200% of the poverty threshold) for multiple group analyses.

The Childhood Executive Functioning Inventory (CHEXI)

The CHEXI consists of 26 items that were developed to represent four subdomains of EF: working memory (11 items, e.g., “Has difficulty remembering lengthy instructions”), planning (4 items, e.g., “Has difficulty with tasks or activities that involve several steps”), inhibition (6 items, e.g., “Has difficulty holding back his/her activity despite being told to do so”), and regulation (5 items, e.g., “Has clear difficulties doing things he/she finds boring”). Parents rated each item using a 5-point Likert rating scale (from 1 = *definitely not true* to 5 = *definitely true*). Higher scores indicate greater EF difficulties.

The Preschool Self-Regulation Assessment (PSRA)

Examiners rated child behavior during testing using the PSRA (Smith-Donald, Raver, Hayes, & Richardson, 2007). Originally designed to complement a battery of self-regulation tasks, the PSRA was used in this study to evaluate children's attentional (e.g., pays attention during instructions), behavioral (e.g., stays in seat) and emotional (e.g., shows pleasure in accomplishment) control during the administration of the *EF Touch* battery. Examiners rated each item on a 3-point Likert rating scale, with each scale point consisting of clear behavioral descriptors. Several items are reverse coded to prevent automatic responding. In line with prior research (Smith-Donald et al., 2007), two subscale scores measuring Attentional/Impulse Control and Positivity were derived. Due to the emotional focus of items comprising the Positivity subscale, the current analyses focus only on the Attentional/Impulse Control subscale.

EF Touch

EF Touch is a computerized battery of EF tasks that was initially created, administered, and extensively studied in paper and pencil (i.e., "flip book") formats (Willoughby & Blair, 2011; Willoughby, Blair, Wirth, & Greenberg, 2010; Willoughby, Wirth, & Blair, 2012). The computerization of tasks has improved the efficiency (i.e., a single data collector with minimal training can now administer the tasks; touchscreen monitors eliminate data keying; data are available in comma delimited format), standardization (i.e., task instructions, examiner prompts, and the administration of training items and resulting decisions about whether a task should be completed are all standardized), and sensitivity (i.e., more difficult items have been added to some tasks; item-level reaction time data is recorded with the aim of differentiating ability level) of the previous battery of tasks (Willoughby & Blair, 2015).

The *EF Touch* program runs in a Windows environment and requires two monitors: a standard monitor displays a script for the interviewer and a capacitive touchscreen monitor records child responses (15" Planar capacitive touchscreen monitors were used). The battery is modular in nature (i.e., any number of tasks can be administered in any desired order) and each EF task takes 3 to 7 minutes to complete. Two warm-up tasks (of 1 to 2 minutes in duration) are typically administered first in order to acclimate children to using the touchscreen (e.g., in a simple reaction-time task, a series of bubbles appears on-screen in random locations and the child is instructed to touch the bubbles as quickly as possible). Because each of the tasks in the *EF Touch* program has been described and extensively studied in detail elsewhere, only abbreviated descriptions are provided here.

Spatial Conflict Arrows. This 36-item spatial conflict task measures inhibitory control. Two buttons appear on the left- and right-most sides of the touchscreen monitor. The child is instructed to touch the button to which the arrow is pointing. Three blocks of 12 arrows are depicted in which arrows either appear above the button to which they are pointing (the congruent condition), above the button opposite the button to which they are pointing (the incongruent condition), or in mixed locations. Each item is presented for 3 s and the accuracy and reaction time of the responses is recorded. The mean accuracy for all of the incongruent items (from the incongruent and mixed conditions) was used to index performance.

Silly Sounds Stroop. This 17-item Stroop-like task measures inhibitory control. Each item displays pictures of a dog and a cat (the left–right placement on-screen varies across trials) and presents the sound of either a dog barking or cat meowing. The child is instructed to touch the picture of the animal that did not make the sound (e.g., touch the cat when a dog bark is heard). Each item is presented for 3 s and the accuracy and reaction time of the responses is recorded. The mean accuracy across all items was used to index task performance.

Animal Go/No-Go. This 40-item go/no-go task measures inhibitory control. Individual pictures of animals are presented, and the child is instructed to touch a centrally located button on the screen every time he or she sees an animal (the "go" response), except when that animal is a pig (the "no-go" response). Each item is presented for 3 s and the accuracy and reaction time of the responses is recorded. The mean accuracy across all no-go responses was used to index task performance.

Working Memory Span. This 18-item span task measures working memory. Each item depicts a picture of one or more houses, each of which contains a picture of an animal, a colored dot, or a colored animal. The child verbally labels the contents of each house. After a brief delay, the house or houses are displayed again without their contents. The child is asked to recall either the animal or the color of the animal that was in each house (i.e., the non-recalled contents serve as a distraction). Items are organized into arrays of 2-, 3-, 4-, and 6-house trails. The mean accuracy of responses was used to index task performance.

Pick the Picture. This 32-item self-ordered pointing task measures working memory. The child is presented with arrays of pictures that vary in length (i.e., 2, 3, 4, or 6 pictures per set). For each set, the child is initially instructed to touch any picture of his or her choice. On subsequent trials within that set, the pictures are presented in different locations and the child is instructed to pick a picture that has not yet been touched. The mean accuracy of responses was used to index task performance.

Farmer. This 36-item spatial visual task (based on Nutley et al., 2010) measures short-term memory. The child is presented with a 4×4 grid of squares that is described as a group of farmer's fields. For each trial, an animal walks through the fields (i.e., appears in successive elements of the grid). After a pause of 1.5 s, the child is instructed to touch the fields through which the animal walked in same order. The items consisted of 2-, 3-, and 4-element sequences. The mean percentage of correctly completed items was used as the analysis variable.

Something's the Same. This 30-item task measures attention shifting and flexible thinking. In the first 20 items, the child is presented with two pictures (animals, flowers, etc.) that are described as being similar with respect to their color, shape or size. A third picture is then presented alongside the original two pictured, and the child is asked to select which of the original pictures is similar to the new picture along some other dimension (e.g., color, shape or size). In the last 10 items, the child is presented with three pictures and asked to identify two of the pictures that are similar and then a second pair from the same three pictures that are similar in a different way to that stated for the first pair. The mean accuracy of responses was used to index task performance.

Analytic Plan

The first research question is to determine whether the CHEXI items are best represented by four (Working Memory, Planning, Inhibition, Regulation) or two (Working Memory & Planning, Inhibition & Regulation) latent factors. Two confirmatory factor analysis (CFA) models were fit to CHEXI items. Because the two models are nested, likelihood ratio tests (LRTs) were used to determine which factor structure best represents the observed data. Global model fit was evaluated using the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the standardized root mean square residual (SRMR) from each model.

The second research question is to determine whether the best-fitting model from the above exhibits measurement invariance based on child gender and household income level. Following the approach described by Widaman and Reise (1997), multiple-group CFA models were estimated to test for configural (i.e., same pattern of factor loadings among groups), weak (i.e., equal factor loadings across groups), and strong (i.e., equal factor loadings and item intercepts) invariance. The fit of these increasingly restrictive models was assessed using both LRTs and CFI difference tests. The LRTs are resported because they represent standard practice. However, given evidence that LRTs have excessive statistical power in large samples (MacCallum, 1990), comparisons of the CFIs are also reported and were more heavily relied on. Following Cheung and Rensvold (2002), models that have CFI differences ≤ .01 are considered to provide an equally good fit to the data.

The third research question is to determine the associations between parent-reported EF behaviors (the CHEXI), examiner-reported EF behaviors (the PSRA), and child performance on an EF battery (*EF Touch*). The best-fitting factor structure derived from part one informs the number of CHEXI subscale scores, which were created by averaging individual items within each subscale. An *EF Touch* composite score was created by averaging task performance across all completed tasks, a strategy that is consistent with other recent work (Willoughby & Blair, 2016; Willoughby, Kuhn, Blair, Samek, & List, in press). Because each task score represents the proportion of correct responses, the composite score has a possible range of 0 to 1. The CHEXI subscale scores were then correlated with the PSRA scores and the *EF Touch* performance.

Item-level descriptive statistics and correlations among the three EF measures (i.e., parent-reported, examiner-reported, performance-based) were computed using STAT v9.3 (2011). All other models were estimated using Mplus v7.31 (Muthén & Muthén, 2015), using full-information robust maximum likelihood (MLR). Because the scaled chi-square statistic that results from MLR cannot be used for likelihood ratio tests in the usual manner, appropriate adjustments were made when comparing nested models (Satorra & Bentler, 2001). Following the guidelines of Hu and Bentler (1999), the following values are considered as being indicative of a good model fit: RMSEA ≤ .05, CFI ≥ .95, SRMR ≤ .08.

Results

Table 1 presents bivariate correlations, means, and *SD*s for the 26 CHEXI items. Individual items have mean scores ranging from 1.65 to 3.32, and all items are significantly correlated (mean r = .38, range = .13–.57, all ps < .001). The observed range of scores for each item is 1 to 5.

Question 1: Factor Structure of the CHEXI

The first research aim was to test whether the four-factor or two-factor model provides the best fit for the data. The fit statistics indicate that both the two- and four-factor models (Models 1 and 2) fit the data reasonably well (Table 2). Although chi-square difference tests indicate that the four-factor model fits better than the two-factor model, $\chi^2(5)$ = 60.54, p < .001, a comparison of the alternative fit indices suggest that both models fit the data nearly identically well (ΔCFI = .00). The parameter estimates from the four-factor model indicate that the correlations between the Inhibition and Regulation (φ = .87) and especially the Working Memory and Planning (φ = .98) latent factors are so large that they are likely redundant. Based on model fit and parsimony, the two-factor model distinguishing Working Memory from Inhibition is deemed to provide the best fit for the observed data.

Before proceeding with tests of measurement invariance for the two-factor model, modification indices were investigated. Modification indices identify two pairs of items that share similar phrasing and have unexplained residual covariation (i.e., items 13 and 18 both reference the child's ability to hold back activity; items 11 and 15 both reference the child's motivation to complete an uninteresting task). The two-factor model was re-estimated allowing for residual covariances between these items (Model 3). The global model fit is good, $\chi^2(296)$ = 686.174, p < .001, RMSEA = .04, CFI = .95, SRMR = .04, and this revised model results in a statistically-significant improvement in model fit (see Model 3 vs. Model 2 in Table 3).

Question 2: Measurement Invariance of the CHEXI

The next set of models tests whether the modified two-factor CFA model from the total sample fits the data equally well for subgroups of children (i.e., gender, household income level). After establishing invariance, group differences in latent means and variances were tested.

Gender

A baseline model that simultaneously fits data for boys (n = 423) and girls (n = 421) without the imposition of any parameter constraints fit the data well, $\chi^2(592)$ = 973.87, p < .001, RMSEA = .04, CFI = .95, SRMR = .04 (i.e., configural invariance, Model 4 of Table 2). Next, all of the factor loadings were equated across groups (i.e., weak invariance, Model 5 of Table 2). This model continued to fit the data well, $\chi^2(616)$ = 1011.61, p < .001, RMSEA = .04, CFI = .94, SRMR = .04, and only resulted in a trivial decrement in model fit relative to the baseline model, $\chi^2(24)$ = 37.57, p = .04, ΔCFI = .01. Moreover, a cross-group examination of factor loadings from the baseline model did not reveal any noteworthy

Table 1. Bivariate Correlations among Individual CHEXI Items.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
CHEXI1	–																									
CHEXI2	.32	–																								
CHEXI3	.47	.17	–																							
CHEXI4	.36	.28	.30	–																						
CHEXI5	.36	.25	.26	.38	–																					
CHEXI6	.55	.26	.39	.39	.45	–																				
CHEXI7	.41	.21	.35	.35	.34	.46	–																			
CHEXI8	.41	.29	.31	.53	.41	.47	.40	–																		
CHEXI9	.46	.22	.47	.33	.35	.54	.35	.46	–																	
CHEXI10	.14	.15	.13	.16	.32	.19	.23	.25	.22	–																
CHEXI11	.38	.30	.32	.46	.38	.40	.39	.50	.36	.30	–															
CHEXI12	.39	.24	.31	.34	.29	.40	.35	.39	.42	.18	.33	–														
CHEXI13	.40	.22	.29	.39	.44	.43	.32	.50	.38	.23	.43	.44	–													
CHEXI14	.51	.31	.42	.42	.35	.53	.45	.46	.54	.19	.39	.50	.46	–												
CHEXI15	.41	.28	.31	.45	.41	.43	.39	.52	.40	.28	.58	.37	.46	.49	–											
CHEXI16	.29	.21	.22	.29	.37	.36	.28	.34	.35	.28	.36	.34	.38	.37	.37	–										
CHEXI17	.45	.16	.43	.30	.31	.42	.38	.30	.38	.14	.31	.42	.34	.44	.36	.34	–									
CHEXI18	.29	.20	.21	.38	.36	.39	.26	.46	.30	.18	.35	.32	.56	.38	.41	.36	.31	–								
CHEXI19	.45	.25	.47	.28	.33	.46	.42	.35	.51	.22	.32	.42	.37	.55	.40	.35	.46	.28	–							
CHEXI20	.57	.34	.46	.37	.41	.59	.49	.47	.55	.21	.44	.49	.49	.66	.50	.41	.49	.37	.63	–						
CHEXI21	.44	.23	.39	.35	.45	.48	.44	.37	.49	.18	.38	.49	.46	.53	.43	.42	.47	.41	.53	.58	–					
CHEXI22	.36	.24	.33	.37	.39	.39	.30	.39	.38	.24	.35	.33	.48	.41	.40	.30	.31	.38	.41	.44	.41	–				
CHEXI23	.48	.24	.36	.32	.33	.41	.41	.38	.42	.21	.37	.35	.36	.47	.38	.34	.41	.29	.46	.52	.44	.37	–			
CHEXI24	.52	.31	.40	.38	.42	.50	.41	.48	.51	.22	.43	.44	.49	.56	.45	.36	.43	.43	.49	.62	.55	.45	.56	–		
CHEXI25	.27	.22	.30	.27	.32	.35	.30	.31	.37	.33	.33	.25	.29	.32	.35	.32	.23	.29	.39	.38	.31	.37	.36	.33	–	
CHEXI26	.40	.24	.35	.28	.35	.42	.38	.37	.45	.23	.36	.39	.38	.44	.39	.32	.45	.32	.51	.48	.51	.44	.45	.46	.39	–
M	2.02	2.62	1.65	2.64	2.74	2.28	2.33	3.07	1.93	3.32	2.76	2.09	2.50	2.18	2.72	2.26	2.00	2.87	1.98	2.14	2.05	2.01	2.12	2.27	2.16	2.06
SD	1.01	1.11	0.80	1.12	1.13	1.00	1.07	1.10	0.91	1.34	1.08	0.95	1.02	1.03	1.10	1.05	1.02	1.16	0.91	0.96	0.91	1.08	1.02	0.95	1.02	1.00

Note. Two participants failed to answer Item 13. The sample size for this item is therefore n = 842 (n = 844 for all other items).

Table 2. Model-Fit Statistics.

Model	Model Description	χ^2	df	*p*	RMSEA	CFI	SRMR
CHEXI Factor Structure							
1	4 Factors (WM, PLAN, INH, REG)	687.80	293	<.0001	.04	.94	.04
2	2 Factors (WM, INH)	752.41	298	<.0001	.04	.94	.04
3	2 Factors (WM, INH) w/correlated errors	686.17	296	<.0001	.04	.95	.04
Measurement Invariance (Gender)							
4	Configural Invariance	973.87	592	<.0001	.04	.95	.04
5	Weak Invariance	1011.60	616	<.0001	.04	.94	.05
6	Strong Invariance	1057.56	640	<.0001	.04	.94	.05
Measurement Invariance (Poverty)							
7	Configural Invariance	993.23	592	<.0001	.04	.94	.04
8	Weak Invariance	1029.30	616	<.0001	.04	.94	.05
9	Strong Invariance	1170.98	640	<.0001	.04	.93	.05

Note. CFI = comparative fit index; INH = inhibition; PLAN = planning; REG = regulation; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; WM = working memory.

Table 3. Nested Model Comparisons.

Nested Tests	$\Delta\ \chi^2$	Δ *df*	*p*	Δ CFI
Model 2 vs. 1	60.54	5	<.001	<.01
Model 3 vs. 2	58.49	2	<.001	.01
Model 5 vs. 4	37.56	24	.040	.01
Model 6 vs. 5	47.77	24	.003	<.01
Model 8 vs. 7	35.57	24	.060	<.01
Model 9 vs. 8	179.47	24	<.001	.01

Note. CFI = comparative fit index.

differences. Finally, all of the item intercepts and factor loadings were equated across groups (i.e., strong invariance, Model 6 of Table 2). This final model fits the data well, χ^2 (640) = 1057.56, $p < .0001$, RMSEA = .04, CFI = .94, SRMR = .04, and the decrement in the model fit is minor, $\chi^2(24) = 47.77$, $p = .003$, ΔCFI < .01. An examination of item intercepts from the previous model did not reveal differences between groups.

Given the evidence for strong measurement invariance for gender, means and variances of the latent CHEXI factors were compared for boys and girls. Following Widaman and Reise (1997), the strong invariance model is parameterized such that the means and variances are scaled to 0 and 1, respectively, for girls (i.e., the reference group) but freely estimated for boys (i.e., the comparison group). Boys have higher scores for both the Working Memory ($\mu = .16$, $p = .04$) and Inhibition ($\mu = 0.21$, $p = .01$) latent factors, which is indicative of more parent-reported difficulties in EF behaviors (Cohen's $d = .15$ and .20, respectively). The latent variance estimates indicate that boys have greater variability than girls for both the Working Memory ($\varphi = 1.27$, $p < .001$) and Inhibition ($\varphi = 1.16$, $p < .001$) latent factors.

Income

Having established strong measurement invariance for gender, the next step was to examine whether the CHEXI model fits equally well for children from high-income ($n = 455$) and low-income ($n = 389$) households. A baseline model that was simultaneously fit to children from high- and low-income households without parameter constraints fit the data well, $\chi^2(592) = 993.23$, $p < .001$, RMSEA = .04, CFI = .94, SRMR = .04 (i.e., configural invariance, Model 7 of Table 2). Next, all factor loadings were equated across groups (i.e., weak invariance, Model 8 of Table 2), resulting in a model that continued to provide a good fit to the data, $\chi^2(616) = 1029.30$, $p < .001$, RMSEA = .04, CFI = .94, SRMR = .05. The decrement in model fit compared to the baseline model is non-significant, $\chi^2(24) = 35.57$, $p = .06$, ΔCFI < .01. Finally, all item intercepts and factor loadings were equated (i.e., strong invariance, Model 9 of Table 2). The resulting model provides an adequate fit to the data, $\chi^2(640) = 1170.98$, $p < .0001$, RMSEA = .04, CFI = .93, SRMR = .04, and the decrement in fit compared to the previous model is trivial, $\chi^2(24) = 179.47$, $p < .001$, ΔCFI = .01. Inspection of intercepts from the previous model did not reveal noteworthy differences across groups.

As with the gender investigations, the latent means and variances were compared for children from high- and low-income households, with high-income households as the reference group. This examination indicates that children from low-income households tend to have higher scores for the Working Memory ($\mu = 0.27$, $p = .001$) latent factor but not the Inhibition ($\mu = 0.11$, $p = .19$) latent factor, indicating greater parent-reported difficulties in behaviors indicative of working memory ability (Cohen's $d = .25$). Compared to their peers from high-income households, children from low-income households also have more variability in the Working Memory ($\varphi = 1.34$, $p < .001$) and Inhibition ($\varphi = 1.43$, $p < .001$) scores.

Question 3: Correlates of CHEXI Subscales

The final goal is to test the association between children's parent-reported EF behaviors (the CHEXI), examiner-reported EF behaviors (the PSRA), and children's EF task performance (*EF Touch*). Partial correlations accounting for gender, race, and household income level do not differ substantively from simple correlations; thus, simple correlations are presented in Table 4. The CHEXI Working Memory subscale is significantly correlated with the Attentional/Impulse Control subscale of the PSRA ($r = -.11$, $p = .001$) and *EF Touch* performance ($r = -.10$, $p = .003$), though these correlations are small. Similarly, the CHEXI Inhibition subscale exhibits a small correlation with the Attentional/Impulse Control subscale of the PSRA ($r = -.11$, $p = .002$) but is not correlated with *EF Touch* performance ($r = -.05$, $p = .19$). *EF Touch* performance is moderately correlated with the Attentional/Impulse Control subscale of the PSRA ($r = .50$, $p < .001$).

Table 4. Bivariate Correlations among EF Measures.

	CHEXI WM	CHEXI INH	PSRA ATT	EF Touch
CHEXI WM	–			
CHEXI INH	0.76	–		
PSRA ATT	−0.11	−0.11	–	
EF Touch	−0.10	$-0.05^{\dagger}$	0.50	–
M	2.08	2.68	2.43	0.65
SD	0.67	0.71	0.64	0.14
Possible Range	1.00–5.00	1.00–5.00	0.00–3.00	0.00–1.00
Observed Range	1.00–5.00	1.00–4.72	0.00–3.00	0.14–0.97

Note. †All correlations significant at $p < .05$ except where marked. ATT = attentional/impulse control; INH = inhibition; WM = working memory.

Discussion

Given the increased interest in measuring EF in young children, the development and evaluation of questionnaire measures is a top research priority. This paper investigates the factor structure, measurement invariance, and correlates of the CHEXI, a questionnaire measure that aims to capture behavioral manifestations of children's EF ability. A two-factor model that distinguishes Working Memory from Inhibition was found to provide the best fit to the observed data and demonstrates strong measurement invariance for subgroups of children (boys vs. girls, high income vs. low income). Anticipated group differences were found. As per Cohen's (1988) conventions, CHEXI scores were found to have a small association with examiner reports of child behavior and EF task performance, whereas examiner-reported behavior has a large association with EF task performance.

To date, all research conducted on the CHEXI has reported that a two-factor solution fits the data well (Catale, Lejeune, et al., 2013; Catale, Meulemans, & Thorell, 2013; Thorell & Nyberg, 2008). The present findings—which are the first to explicitly compare the two- and four-factor models—confirm that a two-factor solution is sufficient. Not only is the fit nearly identical between the two models, but high latent correlations between the Working Memory and Planning factors and the Inhibition and Regulation factors also suggest that the four-factor solution is redundant. This high level of agreement between studies utilizing children of different ages and nationalities supports the continued use of two empirically-distinct subscales (Working Memory and Inhibition) in children aged 3 to 11 years.

With increased interest in questionnaires that assess EF, a corresponding increase in psychometric evaluations of such scales is expected, including formal tests of measurement invariance among subgroups of children. In practice, however, this rarely occurs. The current study is among the first to demonstrate the measurement invariance of an EF rating scale (for another example, see Huizinga & Smidts, 2011) and is the only study to test the measurement invariance of the CHEXI. The present findings of strong measurement invariance suggest that CHEXI factor scores can be directly compared for boys and girls, and for children from high- and low-income households. When comparing factor means between these groups, it was found that being male and low income are risk factors for increased behavioral difficulties in the present sample, although the effect sizes are small (Cohen's d range = .15–.25). These results are consistent with prior research suggesting that boys from low-income households underperform on performance-based measures of EF ability (e.g., Willoughby & Blair, 2015). Gender and household income level are only two of the many groupings that can and should be examined. To the extent that future research will use questionnaire measures of EF behavior among heterogeneous groups of children, including children of different age groups, nationalities, and disability status, greater attention to measurement invariance is needed.

A larger issue in the field of EF measurement concerns the lack of agreement between EF behavioral ratings and task-based performance. Previous studies using the CHEXI have found modest correlations with EF task performance (Catale, Lejeune, et al., 2013; Thorell & Nyberg, 2008), whereas the current study found that only the Working Memory subscale is significantly—albeit weakly—correlated with performance on a battery of EF tasks. Previous reviews (Silver, 2014; Toplak et al., 2013) have suggested that the weak relationship between questionnaire ratings and task performance is partially due to the differing contextual demands and levels of analysis inherent to the two measures. Parental reporting bias is another possible explanation. Previous studies have found that parental and situational characteristics—including depression and parenting stress—may influence parents' perceptions of their children's behavior, including their EF abilities (Joyner, Silver, & Stavinoha, 2009; Silver, 2014).

To better understand the relation between behavioral manifestations and task-based assessments of EF, a third measure is included: examiner reports of child behavior during the EF assessment. It was found that parent reports of behavior (from the CHEXI) correlate weakly with examiner reports of behavior (from the PSRA). However, the examiner reports of behavior are moderately related to child performance on the EF battery. These findings suggest that there are measurable behavioral manifestations of EF that both parents and examiners can report on, but their reports are only weakly related. The weak relation found in the current study may be due to the different respondents, contexts, and questionnaires. Similarly, there are several reasons why examiner-reported behaviors may be more strongly related to child EF performance. First, the child's behavior during the EF assessment has a direct impact on his or her ability to complete the tasks thoroughly and carefully; for example, a child who needs multiple reminders to stay in his or her seat during testing is likely to miss task directions and opportunities to respond to task items. Therefore, ratings of behavior during the assessment situation—as opposed to behavior during non-testing situations—would be expected to correlate more strongly with child performance. The opposite direction of effects is also possible, such that examiners may assign better behavioral ratings to children who demonstrate strong performance on the EF assessment. Disentangling these competing explanations is not possible for the current data set.

Another explanation concerns the extent to which each measure captures executive and non-executive abilities. Similar to how task scores from direct EF assessment may be influenced by non-executive abilities (e.g., reaction time, language), behaviors reported on questionnaires are also likely to be influenced by executive and non-executive processes. Differing degrees of executive and non-executive abilities captured by these three measures likely underlie the differential correlations among them. Although future inquiry into these relations is warranted, the current findings reinforce the notion that behavioral ratings and direct assessment of EF are largely measuring different phenomena. For example, a correlation coefficient of .50 between examiner-reported behaviors and EF task performance indicates that only 25% of the variance in one measure can be explained by the other—and this is likely the best-case scenario, as parent-reported behaviors correlate much less strongly with EF task performance. Therefore, these findings reiterate that questionnaire- and performance-based evaluations of EF should not be used interchangeably, and that further attention needs to be paid to the construct and content validity of these measures.

Theoretical issues aside, it is concluded that the CHEXI is a useful measure for capturing behavioral manifestations of EF in young children. Among existing EF questionnaire measures, the CHEXI certainly fills a unique niche. It is freely available online and is considerably shorter than other rating scales, making it an attractive option for low-resource settings. Further, the factor structure of the CHEXI has now been replicated in several samples—including the present one—and evidence of measurement invariance has been established. Previous work has demonstrated adequate predictive validity as well; parent and teacher ratings from the CHEXI predict academic achievement in children from various European and Asian cultures (Thorell, Veleiro, Siu, & Mohammadi, 2013) and discriminate with good sensitivity and specificity between children with ADHD and typically-developing controls (Catale, Meulemans, & Thorell, 2013; Thorell et al., 2010). Taken together, this body of literature supports the use of the CHEXI with different respondents, cultures, and age groups. However, as norm-referenced scores have not yet been developed, it is impossible to make inferences about whether or not a child's scores are clinically elevated.

It is acknowledged that the conclusions of this study are limited by several factors. First, only parent responses were obtained for the CHEXI, and thus it cannot be confirmed that the factor structure found in this sample would hold for different respondents (i.e., teachers, daycare providers). However, previous work has found that parent- and teacher-derived CHEXI scores function similarly in terms of factor structure (Thorell & Nyberg, 2008). Next, although a diverse group of children from North Carolina and New York was sampled, these findings are not representative of the US as a whole and the mean scores should be viewed as descriptive rather than normative values. Future research utilizing complex sampling methods can provide a better approximation of normative scores for children of different ages. In the present analyses, all 26 CHEXI items are used, although previous studies have excluded items 25 and 26 due to finding a lack of shared variance in them, which may be due to sample characteristics or small sample size. In the current study, these items were not found to be problematic and thus are included. Regardless of the inclusion of these two items, the current findings are consistent with previous work. Future research should examine the ramifications of the use of the 24- or 26-item scale.

Future work should also address the longitudinal measurement invariance of the CHEXI, considering that it is currently recommended for use with children up to ages of 15 years (see http://www.chexi.se). Although the factor structure has been replicated in different age groups (Catale, Lejeune, et al., 2013; Catale, Meulemans, & Thorell, 2013; Thorell & Nyberg, 2008), a formal test of invariance among preschoolers, school-aged children, and early adolescents is warranted. Additionally, the derivation of normed scores for each of these age groups could expand the utility of the CHEXI to a wider range of applications, including the identification of at-risk individuals. Further examination of the CHEXI within typically- and atypically-developing children may also shed light on the clinical usefulness of this scale.

Acknowledgements

The views expressed in this manuscript are those of the authors and do not necessarily represent the opinions and positions of the Institute of Educational Sciences.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Institute of Educational Sciences [grant no.R324A120033].

ORCID

Marie Camerota http://orcid.org/0000-0001-8293-6467

References

Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology*, *28*, 595–616. doi:10.1207/s15326942dn2802_3

Catale, C., Lejeune, C., Merbah, S., & Meulemans, T. (2013). French adaptation of the Childhood Executive Functioning Inventory (CHEXI). *European Journal of Psychological Assessment*, *29*, 149–155. doi:10.1027/1015-5759/a000141

Catale, C., Meulemans, T., & Thorell, L. B. (2013). The Childhood Executive Function Inventory: Confirmatory factor analyses and cross-cultural clinical validity in a sample of 8-to 11-year-old children. *Journal of Attention Disorders*. doi:10.1177/1087054712470971

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 233–255. doi:10.1207/S15328007SEM0902_5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Garon, N., Bryson, S. E., & Smith, I. M. (2008). Executive function in preschoolers: A review using an integrative framework. *Psychological Bulletin*, *134*, 31–60. doi:10.1037/0033-2909.134.1.31

Gioia, G. A., Espy, K. A., & Isquith, P. K. (2003). *Behavior rating inventory of executive function – Preschool version*. Odessa, FL: Psychological Assessment Resources.

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). *Behavior rating inventory of executive function*. Odessa, FL: Psychological Assessment Resources.

Huizinga, M., & Smidts, D. P. (2011). Age-related changes in executive function: A normative study with the Dutch version of the behavior rating inventory of executive function (BRIEF). *Child Neuropsychology*, *17*, 51–66. doi:10.1080/09297049.2010.509715

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. doi:10.1080/10705519909540118

Isquith, P. K., Roth, R. M., & Gioia, G. (2013). Contribution of rating scales to the assessment of executive functions. *Applied Neuropsychology: Child*, *2*, 125–132. doi:10.1080/21622965.2013.748389

Joyner, K. B., Silver, C. H., & Stavinoha, P. L. (2009). Relationship between parenting stress and ratings of executive functioning in children with ADHD. *Journal of Psychoeducational Assessment*, *27*, 452–464. doi:10.1177/0734282909333945

Lohr, S. L. (2009). *Sampling: Design and analysis* (2nd ed.). Boston, MA: Cengage Learning.

MacCallum, R. C. (1990). The need for alternative measures of fit in covariance structure modeling. *Multivariate Behavioral Research*, *25*, 157–162. doi:10.1207/s15327906mbr2502_2

Muthén, L. K., & Muthén, B. O. (2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nutley, S. B., Söderqvist, S., Bryde, S., Humphreys, K., & Klingberg, T. (2010). Measuring working memory capacity with greater precision in the lower capacity ranges. *Developmental Neuropsychology*, *35*, 81–95. doi: 10.1080/87565640903325741.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507–514. doi:10.1007/BF02296192

Silver, C. H. (2014). Sources of data about children's executive functioning: Review and commentary. *Child Neuropsychology*, *20*, 1–13. doi:10.1080/09297049.2012.727793

Smith-Donald, R., Raver, C. C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-regulation Assessment (PSRA) for field-based research. *Early Childhood Research Quarterly*, *22*, 173–187. doi:10.1016/j.ecresq.2007.01.002

STAT Institute Inc. (2011). *SAS/STAT® 9.3 User's Guide*. Cary, NC: SAS Institute Inc.

Thorell, L. B., Eninger, L., Brocki, K. C., & Bohlin, G. (2010). Childhood Executive Function Inventory (CHEXI): A promising measure for identifying young children with ADHD? *Journal of Clinical and Experimental Neuropsychology*, *32*, 38–43. doi:10.1080/13803390902806527

Thorell, L. B., & Nyberg, L. (2008). The Childhood Executive Functioning Inventory (CHEXI): A new rating instrument for parents and teachers. *Developmental Neuropsychology*, *33*, 536–552. doi:10.1080/87565640802101516

Thorell, L. B., Veleiro, A., Siu, A. F., & Mohammadi, H. (2013). Examining the relation between ratings of executive functioning and academic achievement: Findings from a cross-cultural study. *Child Neuropsychology*, *19*, 630–638. doi:10.1080/09297049.2012.727792

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Practitioner review: Do performance-based measures and ratings of executive function assess the same construct? *Journal of Child Psychology and Psychiatry*, *54*, 131–143. doi:10.1111/jcpp.12001

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant & M. Windle (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Willoughby, M. T., & Blair, C. B. (2011). Test-retest reliability of a new executive function battery for use in early childhood. *Child Neuropsychology*, *17*, 564–579. doi:10.1080/09297049.2011.554390

Willoughby, M. T., & Blair, C. B. (2015). Longitudinal measurement of executive function in preschoolers. In J. A. Griffen, P. McCardle, & L. Freund (Eds.), *Executive function in preschool-age children: Integrating measurement, neurodevelopment, and translational research*. Washington, DC: American Psychological Association.

Willoughby, M. T., & Blair, C. B. (2016). Measuring executive function in early childhood: A case for formative measurement. *Psychological Assessment*, *28*, 319–330. doi:10.1037/pas0000152

Willoughby, M. T., Blair, C. B., Wirth, R. J., & Greenberg, M. (2010). The measurement of executive function at age 3 years: Psychometric properties and criterion validity of a new battery of tasks. *Psychological Assessment*, *22*, 306–317. doi:10.1037/a0018708

Willoughby, M. T., Kuhn, L. J., Blair, C. B., Samek, A., & List, J. A. (2016). The test-retest reliability of the latent construct of executive function depends on whether tasks are represented as formative or reflective indicators. *Child Neuropsychology*.Advance online publication. doi: 10.1080/09297049.2016.1205009

Willoughby, M. T., Wirth, R. J., & Blair, C. (2012). Executive function in early childhood: Longitudinal measurement invariance and developmental change. *Psychological Assessment*, *24*, 418–431. doi:10.1037/a0025779