The Factor Structure of Effortful Control and Measurement Invariance Across Ethnicity and Sex in a High-Risk Sample

**Michael J. Sulik · Snjezana Huerta · Argero A. Zerr · Nancy Eisenberg ·
Tracy L. Spinrad · Carlos Valiente · Laura Di Giunta · Armando A. Pina ·
Natalie D. Eggum · Julie Sallquist · Alison Edwards · Anne Kupfer ·
Christopher J. Lonigan · Beth M. Phillips · Shauna B. Wilson ·
Jeanine Clancy-Menchetti · Susan H. Landry · Paul R. Swank · Michael A. Assel ·
Heather B. Taylor**

Abstract Measurement invariance of a one-factor model of effortful control (EC) was tested for 853 low-income preschoolers (M age = 4.48 years). Using a teacher-report questionnaire and seven behavioral measures, configural invariance (same factor structure across groups), metric invariance (same pattern of factor loadings across groups), and partial scalar invariance (mostly the same intercepts across groups) were established across ethnicity (European Americans, African Americans and Hispanics) and across sex. These results suggest that the latent construct of EC behaved in a similar way across ethnic groups and sex, and that comparisons of mean levels of EC are valid across sex and probably valid across ethnicity, especially when larger numbers of tasks are used. The findings also support the use of diverse behavioral measures as indicators of a single latent EC construct.

M. J. Sulik (✉) · S. Huerta · A. A. Zerr · N. Eisenberg ·
T. L. Spinrad · C. Valiente · A. A. Pina · N. D. Eggum ·
J. Sallquist · A. Edwards · A. Kupfer
Arizona State University,
Tempe, AZ, USA
e-mail: msulik@asu.edu

L. Di Giunta
Inter-university Centre for Research in the Genesis
and Development of Prosocial and Antisocial Motivations,
Sapienza University of Rome,
Rome, Italy

C. J. Lonigan · B. M. Phillips · S. B. Wilson ·
J. Clancy-Menchetti · S. H. Landry · P. R. Swank · M. A. Assel ·
H. B. Taylor
The School Readiness Consortium,
Houston, TX 77030, USA

Research on self-regulation has increased dramatically in the past decade and currently is a major topic in developmental science and developmental psychopathology. One reason for this growing interest is that many investigators are finding associations between individual differences in regulatory skills and various aspects of children's socioemotional and academic functioning. For example, various indices of children's regulation have been associated with, and sometimes have prospectively predicted, outcomes in terms of children's social competence (Eisenberg et al. 2001; Spinrad et al. 2006, 2007), sympathy and prosocial behavior (Eisenberg et al. 1996, 1998, 2007b), low levels of externalizing problems (e.g., Kochanska and Knaack 2003; Lengua 2003; Martel et al. 2007; Oldehinkel et al. 2007; Olson et al. 2005; Rydell et al. 2003; Zhou et al. 2008) and internalizing problems (Eisenberg et al. 2007a; Loukas and Roalson 2006; Muris et al. 2004; Silk et al. 2003; compare with Eisenberg et al. 2005b; Murray and Kochanska 2002). In addition, children's regulation has been found to be inversely related to their negative emotionality (Eisenberg et al. 2005b; Rydell et al. 2003) and associated with high quality functioning at school (for reviews, see Eisenberg et al. 2005a; Valiente et al. 2008). Thus, regulatory skills have been integrated into many studies of children's functioning and have predicted a variety of important developmental outcomes.

Scores on various tests of regulatory abilities often are aggregated based on the finding that they tend to relate to one another (e.g., Kochansk et al. 2000). It is possible, however, that different measures of regulation assessing an array of correlated skills would not load on a single latent construct. Thus, one purpose of this study was to examine if low-income preschoolers' scores on a number of measures of regulatory skills load together on a single latent construct (as has been assumed by researchers who use a composite of a number of the same tasks; see Kochanska et al. 2000). In addition, although most of the initial work on children's regulation was conducted with middle-class, primarily European American or European children, measures of regulation are increasingly used in research with children at socioeconomic and other types of risk (e.g., Gilliom 2002; Li-Grining 2007; McCabe et al. (2004); see Raver 2004). Thus, as argued by Raver (2004), it is important to establish measurement invariance for measures of regulation across sociocultural and socioeconomic contexts, including across ethnic/racial and sex groups. Measurement invariance establishes whether a given set of measures taps a particular latent construct such as regulation similarly across various groups so that meaningful inferences across the various groups can be made from data collected with the specific set of measures. Consequently, a second goal of this study was to use confirmatory factor analyses (CFAs) to test the measurement invariance of several measures of regulatory capacities for European American, African American, and Hispanic children in the United States, as well as for boys and girls.

Effortful Regulatory Control: The Construct

Temperament generally is believed to contribute to individual differences in the abilities to self-regulate emotion and behavior. Indeed, effortful control, one of the major components of temperament in Rothbart's (Rothbart et al. 2001) model, is viewed as regulating temperamental, emotional, and behavioral reactivity (Rothbart and Bates 2006). Effortful control (EC) is defined as "the efficiency of executive attention—including the ability to inhibit a dominant response and/or to activate a subdominant response, to plan, and to detect errors" (Rothbart and Bates 2006, p. 129). EC includes skills such as the abilities to shift and focus attention as needed, and to activate and inhibit behavior as needed, especially when one does not feel like doing so. These skills are intimately involved in integrating information, planning, and modulating emotional experience and behavior.

There are numerous measures of EC, but among the most used with young children are Rothbart et al.'s (2001) adult-report temperament scales and Kochanska's et al.'s (1996, 2000) battery of behavioral measures. Rothbart et al.'s Child Behavior Questionnaire contains measures of temperamentally based attention focusing and inhibitory control (the ability to inhibit behavior when one is motivated to act). These two capacities are interrelated and tend to load together on a single factor (Eisenberg et al. 2004; Rothbart et al. 2001).

The typical behavioral measures of EC tap children's abilities to delay, to suppress or initiate behavior, to focus attention and persist, and to execute gross or fine motor control (e.g., walk slowly; e.g., Kochanska et al. 2000; Murray and Kochanska 2002). Sometimes measures of EC quite explicitly tap executive attention (Carlson 2005; Riggs et al. 2006; also Kochanska et al.'s 2000, adapted Stroop task), such as the ability of children to knock on a table (i.e., closed fist) when they see an experimenter tap on a table (i.e., open palm) and vice versa. Given the diverse abilities required to successfully perform across a variety of these tasks, it is reasonable to question if they all load on a single construct. However, if all tasks in a set tap primarily effortful control, they might load on a single latent construct despite some differences in the context in which regulatory capacities are expressed and measured.

There are conceptual reasons to hypothesize that the tasks typically used to assess EC might load on more than one factor, despite intercorrelations among these measures. Eisenberg and colleagues (e.g., Eisenberg and Morris 2002; see also Carver 2005; Derryberry and Rothbart 1997; Nigg 2000) have attempted to differentiate the truly effortful and voluntary self-regulatory processes involved in EC from other aspects of control or constraint (or the lack thereof) that seem to be involuntary or so automatic that they often are not under voluntary control. These *reactive control* (RC) processes refer to relatively involuntary motivational approach and avoidance systems of response reactivity that, at extreme levels, result in impulsive undercontrol and rigid overcontrol. Measures of RC typically assess impulsivity (speed of response initiation, including surgent approach behaviors and approach to attractive objects) and overcontrol (rigid, constrained behavior) or behavioral inhibition (slow or inhibited approach, distress or subdued affect in situations involving novelty or uncertainty; Derryberry and Rothbart 1997; Kagan and Fox 2006). Pickering and Gray (1999) and others have argued that the approach and avoidance motivational systems related to impulsive and overly inhibited behaviors, respectively, are associated with subcortical systems such as Gray's Behavioral Activation System (BAS; involving sensitivity to cues of reward or cessation of punishment) and Behavioral Inhibition System (BIS; activated in situations involving novelty and stimuli signaling punishment or frustrative nonreward).

Consistent with the distinction between EC and RC, adults' reports of children's reactive overcontrol and

undercontrol tend to load on a different factor than does EC in confirmatory factor analyses (Eisenberg et al. 2004; Rothbart et al. 2001; Valiente et al. 2003). Thus, it is quite possible that some behavioral tasks used to assess RC might not load with other measures of EC in CFAs. Tasks involving rewards or punishment (e.g., prizes or loss of points) may measure impulsivity (i.e., approach to rewards) as much or more than EC (e.g., modulation of attention or inhibition of the urge to approach). It is also possible that such tasks tap primarily EC for some children (especially those who behave in a regulated manner), a preponderance of reactive BIS/BAS responding for other children, or some of both for other children. Moreover, on adult-report measures of EC, respondents may have some difficulty differentiating EC from low impulsivity, so reports of children's EC could partly reflect children's levels of impulsivity or behavioral overcontrol. In addition, tasks involving the simple delay of a prepotent or automatic response may load on a different factor than tasks that assess inhibition when one has to hold a rule in mind, respond according to the rule, and inhibit a prepotent response (as in the knock/tap task). In fact, Garon et al. (2008) found that the ability to perform the former types of tasks generally develops prior to the ability to perform well on the latter tasks.

Empirical Findings on the Structure of the Construct of Effortful Control

Some researchers have found that in North American, predominantly middle-class and European American samples, various behavioral measures of EC are intercorrelated (e.g., Kochanska et al. 2000) and the reliability for a battery of EC tasks is high for children aged 33 and 42 months (Kochanska and Knaack 2003), However, using principal components analyses with a battery of EC tasks, Murray and Kochanska (2002) obtained multiple groupings or components. For toddlers, they found two components (using 6 tasks): one for delay and gross motor movement and one for the abilities to suppress or initiate behavior. For a sample of preschoolers, they obtained four components (using 13 tasks): delay, gross motor control, fine motor control, and suppress/initiate behavior. For children in the early school years, they found two components (using 7 tasks): motor control and suppress/imitate (they did not have delay tasks). If the measures of EC involve diverse skills, they may not load on the same construct.

Other investigators using factor analyses also have found that tasks assessing reactive undercontrol (i.e., impulsivity, approach to reward) sometimes load on a different factor than tasks that appear to more clearly assess EC. Kindlon et al. (1995) assessed a number of behavioral measures of "impulsivity" and found two clusters of measures: (a) an

inhibitory control factor (reflecting responses to a Stroop task, the number of times the child required redirecting back to tasks, the ability to stop behavior in response to a signal, and the ability to inhibit a strong competing response on the trail-making task) and (b) a factor believed to reflect insensitivity to punishment or nonreward, that is, BAS/BIS types of responding (e.g., the relative failure to exhibit nonresponse within a motivationally salient context of earning money or points). Olson et al. (1999) also obtained two factors with somewhat similar tasks: one that seemed to reflect inhibitory control (e.g., the ability to inhibit motor behavior on command in walking and drawing tasks, as well as reflective performance on the Matching Familiar Figures Task [MFT]), and one that reflected the ability to wait patiently for a reward (i.e., a gift). However, Olson et al. (1990) found that whereas measures of inhibitory control (MFT performance, motor inhibition) tended to load on one factor, rated task orientation (which likely partly reflected inhibitory control) and delay of gratification (attempts to open a gift prematurely) tended to load on the other.

Yet another group of investigators obtained different factors for behavioral and reported measures of regulation/control. White et al. (1994) found that older children's delay of gratification (the ability to inhibit playing a game in the face of losing rewards) and other measures believed to reflect inhibitory control [Stroop errors, circle tracing]) all loaded together when factored with other-report measures of undercontrol, motor restlessness, impulsivity, and lack of persistence, and self-reported impulsivity (all of which loaded on a second factor).

The lack of consistency across studies is understandable because these studies differed not only in the tasks used and in the number of tasks, but also in the age range of the children involved and sometimes in regard to sex, risk status, or ethnicity. For example, unlike the other studies discussed, White et al. (1994) included only boys and was over 50% African American. Nonetheless, some of the aforementioned findings suggest that tasks involving rewards or attractive objects might elicit reactive undercontrol and load on a separate latent construct than tasks that tap only the abilities to inhibit and/or activate behavior as required.

In the present study, we administered seven EC tasks to a relatively large sample of children from low-income families. Some of the tasks assessed would seem to be relatively pure measures of effortful control (the knock/tap task used as an index of executive functioning and tasks involving persistence on a boring task or activation and inhibition of behavior upon command) whereas other tasks may have assessed reactive undercontrol in addition to EC, at least for some children (e.g., tasks involving waiting for a gift). One task also involved motor skills similar to the task

that loaded on Murray and Kochanska's (2002) factor reflecting preschoolers' gross motor skills but on the suppress/initiate grouping for young school children. Unlike in the aforementioned studies, we used confirmatory factor analyses as well as exploratory factor analyses to examine if a battery of EC indices could be considered as measuring one construct. With CFAs, one can explicitly test if a given set of measures load on a single conceptually based latent construct. Based on prior research, and because our sample was primarily low-income whereas most other studies involved more diverse or middle-class samples, we were unsure whether or not the behavioral tasks would load on one primary factor.

We also obtained teachers' reports of children's EC. In prior work, such reports have tended to load with behavioral measures on a single latent construct (e.g., Eisenberg et al. 2004; Spinrad et al. 2007). Thus, we predicted that teacher-reported EC would load with the behavioral measures of EC in both exploratory factor analyses and a CFA.

Variations in Regulatory Skills as a Function of Children's Ethnicity and Sex

Much of the work on EC has been conducted with primarily European-American samples. However, some research suggests that many measures of EC also can be used with minority and low-income children. For example, McCabe et al. (2004) found reasonable variation in some tasks with a disadvantaged, primarily Hispanic and African-American sample. Li-Grining (2007), with a sample of low income, mostly minority children in three United States cities, found the expected relations of regulatory abilities with age and with risk factors.

Within the United States, there is some, albeit limited, evidence that minority children are at risk for poor regulation abilities in comparison to European Americans. For instance, Hispanic 5th and 6th graders reported lower levels of EC than their European American peers (Loukas and Roalson 2006). In another study, Hispanic children were marginally significantly higher than European American children on executive control tasks, but did not differ in delay of gratification tasks (Li-Grining 2007). In a recent study, Aikens et al. (2008) reported higher levels of attention problems in African American than for Hispanic and European American children.

Consistent with the possibility of ethnic/racial differences in effortful regulation, there may be some ethnic differences in negative emotionality among minority groups. African American toddlers have been observed to be more negative towards their mothers than European American and Mexican American children; in addition, less acculturated Mexican American toddlers were seen as lower in negativity than European American children (Ispa et al. 2004).

If there are differences in regulation in different ethnic groups, they may be at least partially accounted for by disparities in socioeconomic status and associated risks. Indeed, poverty has been negatively related to regulation abilities in a number of studies (Evans and English 2002; Howse et al. 2003; Li-Grining 2007; Mezzacappa 2004; Noble et al. 2005). Of course, distinct cultural values and parenting beliefs also may account for differences in children's regulation.

A somewhat different question is whether or not various indices of EC are invariant in their loadings and in their intercepts on the construct of EC. Despite Raver's (2004) arguments regarding the importance of examining the measurement invariance of children's regulation measures across ethnic/racial groups, to our knowledge, this issue has not been addressed. To evaluate measurement invariance, CFAs are typically used to test for configural invariance (if the same factor structure can be specified in each group), metric invariance (if the same factor loadings for items can be specified in each group), and scalar invariance (if the same factor loadings *and* intercepts for like items' regressions on the latent variables can be specified in each group). These tests are important because without evidence of measurement invariance, it is difficult to determine whether cross-ethnic variations are due to error or measurement artifact rather than to other factors.

There is an emerging body of literature suggesting that high EC or self-regulation in low-income, minority populations predicts low levels of problem behaviors (Fantuzzo et al. 2001; Gilliom et al. 2002; Miller et al. 2006), replicating findings using White, middle-class samples. Moreover, Li-Grining (2007) found no evidence of ethnic/racial differences in the relations of measures of children's delay of gratification and executive control to negative emotionality (or to risk factors or mother–child interaction). These findings provide preliminary evidence of the predictive validity of measures of regulation (including EC) with low-income, minority children (Raver 2004; Mendez et al. 2002) and indicate that measures of regulation may function similarly for children differing in socioeconomic status or ethnicity. Based on this pattern of findings, we expected to find measurement invariance across groups.

In terms of measurement invariance of EC/regulation across sex, investigators have generally found that the measurement of regulation is similar for boys and girls. For instance, Windle (1992) computed CFAs and found the Revised Dimensions of Temperament Survey (Windle and Lerner 1986), which includes an attentional focusing subscale (a component of EC), to be equivalent across sex. Similarly, Kim et al. (2003) found that the same components of temperament (including a subscale of attention) were evident for both boys and girls on the Early

Adolescent Temperament Questionnaire (Capaldi and Rothbart 1992). It should be noted, however, that researchers examining measurement invariance in EC/regulation have not relied on behavioral measures; thus, it is important to examine whether observational measures of EC hold together similarly for boys and girls. Based on the existing studies, as well as on the fact that relations of EC to variables such as adjustment generally have not been moderated by sex (e.g., Eisenberg et al. 2004; Eisenberg et al. 2005b), we expected to find invariance across sex, especially in terms of the degree to which the various indices loaded one or more constructs of EC (configural and metric invariance).

There is evidence suggesting that mean levels of EC might vary with the sex of the child. Girls are generally viewed as more regulated than are boys. Else-Quest et al. (2006) reported meta-analyses of sex differences in 3 month to 13-year-old children's temperament involving 205 studies yielding a total of 1,758 effect sizes. Although most temperament dimensions did not show clear sex effects (e.g., emotionality), there was a large effect size for effortful control favoring girls. However, the data were based mostly on parents or teachers' reports, and only a few of the included studies involved behavioral observations of EC. There is some evidence that girls outperform boys on behavioral measure of EC (Blair et al. 2005; Kochanska et al. 1996, 2000; Olson et al. 2005), although gender differences on behavioral measures of EC tend to be smaller and less consistent than for adult-report measures. Due to differences in gender roles, some tasks may have differential appeal or significance for boys and girls; moreover, gender stereotypes could affect adults' ratings of children's EC. Because investigators often compare mean levels of boys' and girls' EC/self-regulation, it is important to examine whether there is scalar invariance across sex (which justifies such comparisons).

Methods

Demographics

Participants were drawn from 53 preschools in and around Houston, Texas and 58 preschools in and around Tallahassee, Florida, for an intervention project (these data are from baseline). To be included in the study, at least 60% of students at each center had to be eligible for free or reduced lunch. Potentially eligible preschools were identified through Head Start directors and independent school districts in Texas, and through the website for the Florida Department of Children and Families in Florida; preschool directors provided information about free/reduced lunch rates. All eligible preschools that agreed to participate in the study were included in this sample. Usually only one

classroom at a preschool met eligibility criteria, but if multiple classrooms were eligible, a specific classroom was selected for participation based on recommendations from directors and agency leaders. If more than eight children had permission to participate in a given classroom, eight were randomly chosen. In Texas, only one class was used per school; in Florida, two classes were occasionally used because some preschool directors did not want the curriculum to differ across classrooms. The data were collected across two school years; different schools were involved each year. The number of male and female participants was comparable across sites: In Texas, there were 197 males and 221 females, and in Florida, there were 212 males and 223 females. There were approximately equal numbers of African American students across the sites. European American students, however, were almost exclusively located at the Florida site, whereas Hispanic students were almost exclusively located at the Texas site. In Texas, the sample consisted of five European American students, 225 African American students, and 188 Hispanic students. In Florida, the sample consisted of 224 European American students, 197 African American students, and 14 Hispanic/Latino students (likely of Cuban origin). Mean age at testing was 4.52 ($SD = .40$) in Texas and 4.45 ($SD = .50$) in Florida. This difference was statistically significant, $t(853) = 2.34$, $p<.05$, $r^2 = .01$, but too small to be of practical importance. The primary caregiver, usually the mother, reported educational attainment was reported on a 10-point scale: 1 = middle school, 2 = some high school, 3 = high school diploma, 4 = vocational training, 5 = some college, 6 = associates degree, 7 = bachelor's degree, 8 = graduate school but no degree, 9 = masters degree, 10 = doctorate. The mean level of maternal education was 3.62 ($SD = 1.67$) in Texas and 4.44 ($SD = 1.74$) in Florida. Although the level of educational attainment was low at both sites, it was higher in Florida than in Texas, $t(590) = 5.86$, $p<.01$, $r^2 = .05$. An ANOVA was run to examine whether the level of maternal education differed across racial/ethnic groups. This test was significant, $F(1,590) = 36.92$, $p<.01$, $r^2 = .11$, so post-hoc tests were run using the Tukey adjustment to maintain a .05 type I error rate. These comparisons indicated that European Americans ($M = 4.41$) and African Americans ($M = 4.33$) did not differ in maternal education, but that Hispanics ($M = 3.03$) were significantly lower than European Americans and African Americans on this variable.

Behavioral Measures of Effortful Control

Children from sites in Texas and Florida were tested by a team of experimenters at their preschools in early fall of 2006 or 2007. University personnel and staff members drawn from the community (about 65%) served as

experimenters. There were 39 experimenters in Texas (82% female) and 57 in Florida (67% female). In Texas, 21% of experimenters were European American, 38% were Hispanic, 28% were African American and 13% were other ethnicities; in Florida, 84% were European American and 16% were African American. Experimenters were trained by expert staff and allowed to practice until they felt comfortable administering each task. Experimenters were required to be certified by demonstrating consistent and accurate administration on every task before data collection started, and were monitored during the first week and intermittently throughout data collection to ensure quality. After summer breaks, experimenters were retrained to minimize drift.

Bilingual experimenters were available if needed for Spanish speaking children. Parents who indicated in the consent packet that their children had exposure to spoken Spanish received a follow-up phone call to assess home language use. If parents reported that the child used Spanish more than 50% of the time, the assessment was administered in Spanish; otherwise, the assessment was administered in English (although Spanish was used when deemed necessary by the experimenter). Bilingual and Spanish speaking children were found to have difficulty understanding words used to label body parts for the bird and dragon task. This was due to the many different terms used by this subgroup to label these body parts. Therefore, examiners asked the children to label various body parts using their own words. The labels that the children provided were used during the procedure to ensure understanding for this task. Half (52%; 106 children) of Hispanic children had an assessment partially or fully in Spanish.

With rare exceptions, the six videotaped behavioral tasks were administered in one session and in a constant order. A computer-administered continuous performance task (CPT) was usually administered during a different session.

The behavioral measures were scored by a main coder, as well as by a reliability coder who scored 24% to 32% of the data, depending on the task. Quality of implementation was coded dichotomously by both the main and reliability coders as usable or not usable. Inter-rater agreement on quality of implementation ranged from 92% to 100%. Because agreement between the two coders was high, the main coder's determination of quality of implementation was used. Intraclass correlation coefficients (ICCs) for the children's performance on the following measures ranged from .92 to .99, and are reported individually for each task below. The reliabilities were nearly identical when computed for all cases and when computed only for instances in which the data were scored as usable, so only the latter reliabilities are presented.

Knock tap During the knock tap task, an experimenter either tapped on the table with an open, flat hand or knocked on the table with a closed fist (Luria 1966; Perner and Lang 2000). For the first eight trials of this task, the child was asked to imitate the experimenter, knocking when the experimenter knocked and tapping when the experimenter tapped. Then, for the subsequent eight trials, the child was instructed to reverse his or her actions (i.e., to knock when the experimenter tapped and to tap when the experimenter knocked). The proportion of correct responses during the eight reversed trials was calculated (*ICC* = .99).

Rabbit turtle During the rabbit turtle task, the child was asked to maneuver a turtle (slowly) and a rabbit (quickly) to follow a curved path from one end of a mat to a toy barn at the other end (Kochanska et al. 2000). First, the experimenter demonstrated how to travel the path to the barn with a boy or girl figure, and the child completed two baseline trials with this (same-sex) figure. Next, the experimenter presented a rabbit figure, who wanted to travel on the path to the barn and was the "fastest rabbit in the world". Thus, the task was to move the rabbit to the barn quickly while still staying on the path. The child then completed two timed trials with this rabbit figure. Last, the experimenter explained that the "slowest turtle in the world" also needed to travel the path to the barn, requiring the child to slow down his/her behavior to travel the path as slowly as possible. The child then completed two timed trials with this turtle figure. The difference between the average time for rabbit trials and the average time for turtle trials was calculated and divided by 60 to obtain a single score of time in minutes (*ICC* = .93).

Yarn tangle During this task, the child was instructed to untangle a ball of yarn while the experimenter left the room for 2 min (Goldsmith et al. 1993). Coders assessed the child's persistence to untangle the yarn for every 10 s interval (1 = child did not attempt or actively refused to engage in the task; 2 = child was minimally engaged in the task, either briefly or sporadically, showing little effort; 3 = child was engaged in the task for about half of the interval and then quit, or worked at it on and off; 4 = child was engaged in the task for most to all of the interval; 5 = child was intensely and actively engaged in the task for most to all of the interval, and never actively quit task). These interval scores were averaged across the task to obtain an average persistence score (*ICC* = .94).

Gift wrap The gift wrap task required the children to remain seated and face forward for one minute as a gift was noisily wrapped behind them while the experimenter instructed the child not to "peek" (Kochanska et al. 2000). The latency to peek was calculated as the number of seconds elapsed from when the experimenter gave the directions and commenced rustling the tissue paper to when

the child attempted to peek or the end of the minute, whichever came first. The latencies were divided by 60 s to obtain a score of elapsed time as a proportion of the one-minute maximum (ICC = .96).

Waiting for bow In the waiting for bow task, a wrapped gift was placed on the table within the children's reach while the experimenter explained that the bow that was meant to go on the gift was forgotten (Kochanska et al. 2000). The children were asked to stay in their seat and not to touch or open the gift while the experimenter stepped outside of the room to retrieve the bow. The gift box was a small box with a removable lid; both the base of the box and the lid were wrapped in brightly colored paper. Inside the box was a small gift nestled in the folds of some tissue paper. The gifts, which varied and consisted of stickers, bracelets, spin tops, rings, sunglasses, or toy frogs, were given to the children as a reward for their participation.

Waiting for bow lasted two minutes. Three latency times were measured: the latency to touch the box, the latency to peek inside the box, and the latency to extract the contents of the gift box. The latency to touch was the number of seconds elapsed from when the experimenter had given the instructions and walked away to when the child touched the box or the end of the two minutes, whichever came first. Touches consisted of any intentional contact with the gift. Similarly, the latency to peek was the number of seconds elapsed before the child either lifted or removed the lid in order to look inside the gift box, or the maximum of the two minutes. The latency to remove the gift's contents was the number of seconds elapsed before the child removed the toy itself or the tissue paper in which it sat, to a two-minute maximum. The three latencies were then averaged to obtain a single score (ICC = .92). The latency composite was not calculated if the child did not spend at least one minute with the gift while the experimenter was gone; 68 children had missing data for this task because they went away from the gift without first taking it out.

Bird and dragon The bird and dragon task tested both inhibitory and activational control. Much like the game "Simon-says", in this task, the children were asked to perform the commands issued by a bird puppet, which the experimenter described as the nice puppet, and not to comply with the commands issued by the dragon puppet, which the experimenter described as being mean (Reed et al. 1984; Kochanska et al. 1996). The experimenter followed a scripted command pattern that included five bird commands and seven dragon commands. Children's responses to each bird command were assessed on a scale of 0 (no movement) to 3 (full, correct movement) and were reverse scored for the dragon commands. Inhibition and activation scores, for the dragon and bird commands

respectively, were calculated as the average performance for the commands of each kind.

Upon examination of this task in another ongoing study, it became clear that there is a potential problem with the standard method of scoring this measure. Children's inhibition during the dragon's commands could be due to effortful, inhibitory control or, alternatively, due to general inhibition or lack of cooperation that resulted in their not performing any or most of the commands (including the bird's activation commands). If children enacted none of the commands during the entire task, they would receive a perfect score for inhibition (and a very low score for activation) simply because they were inhibited or not cooperating. To address this problem, if children did not respond to at least two of the five bird trials, their inhibition scores were set to missing (this was done for 38 children).

Continuous Performance Task (CPT)

Using a shortened and adapted version of the original CPT (Rosvold et al. 1956), the child was seated at a computer and instructed to press the space button as soon as the image of the target stimulus (i.e., a fish) appeared on the screen. One-hundred fifty dot-matrix pictures of ten different familiar objects (e.g., butterfly, flower) were randomly presented on the screen, including 30 presentations of the target stimulus and 120 presentations of non-target stimuli. Each stimulus appeared on the screen for 0.5 s with 1.5 s intervals between stimuli. The proportion of errors of omission (i.e., when the child failed to press the button in response to the presentation of the target stimulus) was calculated. The original scores, which were the proportion of errors of omission, were subtracted from 1 to get the proportion of trials without an error of omission. Accordingly, high scores reflect high attentional control. Fifteen children did not receive a score on the CPT because they completed fewer than 75% of the trails; in addition, children in Florida during the first year received the wrong computer game and could not be used (n = 638 for the analyses).

Teacher Questionnaires

One to 2 months after the above described tasks were administered, the 13-item attention focusing and 14-item inhibitory control scales of the Child Behavior Questionnaire (CBQ; Rothbart et al. 2001) were distributed to the children's teachers. Teachers rated children on a seven-point scale (1 = never, 7 = always); 827 teachers completed these measures. Reliability for both scales was good (both αs = .86). Because these two scales were highly related, $r(825) = .82$, they were subsequently averaged to create a single score.

Missing Data

The primary reasons for missing data on the behavioral tasks were poor quality of implementation (e.g., the experimenter deviated from the script) and technical difficulties (e.g., poor audio or visual quality). For each task, less than 5% of the sample was excluded due to poor quality of implementation and less than 4% of the sample was excluded due to technical difficulties. Rates of complete data are as follows: bird and dragon, 87.7%, $n = 748$; gift wrap, 89.8%, $n = 766$; yarn tangle, 90.2%, $n = 769$; knock tap, 90.2%, $n = 769$, rabbit turtle 92.5%, $n = 789$; and waiting for bow, 83.2%, $n = 710$.

Analyses were conducted to verify that ethnicity and sex were unrelated to missingness on the eight variables used in our factor analytic models. First, three one-factor MANOVAs predicting missingness on all eight measured variables were run for the entire sample using sex as a predictor and for each site separately using ethnicity as a predictor (because site differences in missing data could be confounded with ethnic differences). The MANOVAs for sex in the entire sample, $F(8, 825) = 1.42$, and for ethnicity in Florida, $F(8, 412) = 1.72$, were nonsignificant, but the MANOVA for ethnicity in Texas was significant, $F(8, 404) = 4.21$, $p<.001$.

Follow-up ANOVAs were run to determine which variables showed ethnic differences in missing data rates in Texas. African Americans had higher rates of missing data than Hispanics for knock tap, $F(1, 411) = 4.50$, $p<.05$, $r^2 = .01$, gift wrap, $F(1, 411) = 5.65$, $p<.05$, $r^2 = .01$, waiting for bow, $F(1, 411) = 8.49$, $p<.01$, $r^2 = .02$, and the CPT, $F(1, 411) = 6.58$, $p<.05$, $r^2 = .02$. Hispanics had higher rates of missing data than African Americans for bird and dragon, $F(1, 411) = 5.83$, $p<.05$, $r^2 = .01$. There were no group differences in performance for any of these tasks, however, suggesting that ethnicity was unrelated to the values of the missing data and did not bias the results.

Results

Data Preparation

One variable, rabbit turtle time, had two extreme outliers. These outliers were replaced with a value slightly larger than the next smallest value for this variable (Tabachnik and Fidell 2006). Prior to conducting the analyses, transformations were applied to correct for skewness that could potentially bias the analysis. Following Tabachnik and Fidell's (2006) recommendations, a square root transformation was applied to the rabbit turtle time to correct for moderate negative skewness. Descriptive statistics for the variables prior to transformation are displayed in Table 1.

Correlations among the measures of EC are displayed in Table 2.

Data Analytic Strategy

Three sets of analyses were run using Mplus 5.1. Mplus uses maximum likelihood as a missing data treatment (Muthén and Muthén 2007), which produces unbiased parameter estimates when data are missing at random (Schafer and Graham 2002). First, exploratory factor analyses (EFAs) were run on each ethnic group separately, and for boys and girls separately (collapsing across ethnic groups), to confirm that all indicators loaded on a single factor, and that all indicators had substantial loadings on that factor. In these models, one- and two-factor solutions were estimated with eight observed indicators, which included the seven behavioral measures, and teachers' reports of effortful control from the CBQ. Next, confirmatory factor analyses (CFAs) were run separately for each ethnic group and separately by sex using a one-factor solution with the same indicators. After verifying that the CFAs had good fit in each group, multi-group CFAs were run to establish factorial invariance across the three ethnic groups and across the sexes.[1]

Because the value of p associated with the χ^2 statistic is related to sample size (Kline 1998), we relied upon alternative goodness of fit indices that are less sensitive to sample size. These included the Root-Mean-Square Error of Approximation (RMSEA) and the Standardized Root-Mean-Square Residual (SRMR). For the RMSEA, values less than .05 are small, and values between .05 and .08 are acceptable (Browne and Cudek 1993); SRMR values greater than .08 are indicative of relatively poor fit (Kelloway 1998). No incremental fit indices (e.g., CFI) are reported because the standard null model estimated and utilized in generating incremental fit indices is not an appropriate comparison in multiple-group factor analysis models (Widaman and Thompson 2003). To test measurement invariance, χ^2 difference tests were performed to compare nested models (Byrne 1994; Kline 1998), adopting an alpha level of .05.

---

[1] Garon et al. (2008) distinguished between simple and complex inhibition, and the two-factor EFAs suggested that the three delay/persistence tasks (gift wrap, waiting for bow, and yarn tangle) primarily loaded on one factor, whereas the other tasks loaded on a second factor, providing some support for this distinction in our data despite the small second eigenvalues. In CFAs, the Bayesian Information Criterion (BIC; Raftery 1995) did not clearly favor one model over the other, and the correlations among the factors in the two-factor model were estimated to be .75 or above in each group. Moreover, when two factors were tested in each of the ethnic groups, the model would not converge for the Hispanic group. Therefore, we used the one-factor model to maintain parsimony.

Table 1 Descriptive statistics

Variable	N	Mean	SD	N	Mean	SD	N	Mean	SD
	Entire Sample			Male			Female		
Knock Tap[a]	755	0.59	0.36	360	0.56	0.37	395	0.61	0.35
CPT[b]	625	0.93	0.05	293	0.93	0.05	332	0.94	0.05
Bird/Dragon[c]	687	3.00	1.18	335	2.90	1.19	352	3.11	1.15
Gift Wrap[d]	751	0.57	0.41	358	0.51	0.41	393	0.63	0.41
Waiting for Bow[d]	696	1.66	0.46	323	1.67	0.46	373	1.66	0.46
Yarn Tangle[e]	754	3.31	1.55	358	3.15	1.55	396	3.46	1.54
Rabbit Turtle[f]	774	1.92	5.10	372	2.40	5.77	402	1.47	4.34
Teacher CBQ[g]	775	4.47	0.94	372	4.27	0.91	403	4.66	0.92
	European American			African American			Hispanic		
Knock Tap	200	0.62	0.36	374	0.47	0.36	181	0.58	0.36
CPT	158	0.94	0.05	298	0.93	0.05	169	0.93	0.05
Bird/Dragon	192	3.16	1.14	349	2.85	1.23	146	3.18	1.03
Gift Wrap	205	0.72	0.39	368	0.53	0.41	178	0.50	0.40
Waiting for Bow	180	1.81	0.31	339	1.60	0.50	177	1.65	0.47
Yarn Tangle	202	3.82	1.34	375	3.34	1.53	177	2.69	1.61
Rabbit Turtle	207	3.71	6.75	386	1.11	3.81	181	1.60	4.82
Teacher CBQ	203	4.49	1.05	392	4.40	0.90	180	4.61	0.86

Descriptive statistics are for variables prior to transformation

[a] Proportion of trials correct

[b] Proportion of trials without an error of omission

[c] Inhibition score for dragon trials (range: 1–4)

[d] Minutes

[e] Persistence score (range: 1–5)

[f] Seconds

[g] Average of inhibitory control and attention focusing scales on the CBQ (range: 1–7)

Exploratory Factor Analyses

EFAs were run separately for each ethnic group and separately for males and females to determine how many factors should be used and to determine whether each indicator loaded substantially on at least one factor. For each group, a one-factor solution and a two-factor Geomin rotated solution were extracted. Scree plots revealed that each EFA produced only one large eigenvalue, suggesting that a one-factor solution was most appropriate (eigenval-

Table 2 Correlations among measures of effortful control

		1	2	3	4	5	6	7	8
1	Knock Tap	–							
2	CPT	0.22	–						
3	Bird/Dragon	0.39	0.28	–					
4	Gift Wrap	0.27	0.22	0.35	–				
5	Waiting for Bow	0.09	0.15	0.14	0.28	–			
6	Yarn Tangle	0.18	0.14	0.21	0.32	0.19	–		
7	Rabbit Turtle	0.17	0.11	0.25	0.19	0.08	0.17	–	
8	Teacher CBQ	0.19	0.23	0.29	0.26	0.22	0.17	0.15	–

All correlations significant at $p < .05$

ues were 2.40 and 1.06 for males, 2.60 and .99 for females, 2.74 and 1.06 for European Americans, 2.47 and 1.15 for African Americans, and 2.34 and 1.18 for Hispanics).[2]

Confirmatory Factor Analyses

Confirmatory factor analyses were run separately for boys and girls and separately for each ethnic group. The CFAs for males, $\chi^2(20, n = 409) = 30.15$, $p = .07$, $RMSEA = .035$ (90% $CI = .000–.060$), $SRMR = .037$, and females $\chi^2(20, n = 444) = 42.28$, $p<.01$, $RMSEA = .050$ (90% $CI = .029–.071$), $SRMR = .041$, both fit the data well. Likewise, the CFAs for the European American group, $\chi^2(20, N = 229) = 15.19$, $p = .76$, $RMSEA = .000$ (90% $CI = .000–.040$), $SRMR = .037$, African American group, $\chi^2(20, N = 422) = 46.89$, $p<.01$, $RMSEA = .056$ (90% $CI = .036–.078$), $SRMR = .047$, and Hispanic group, $\chi^2(20, N = 202) = 24.42$, $p = .22$ $RMSEA = .033$ (90% $CI = .000–.060$), $SRMR=.048$, all fit the data at least acceptably well. The unstandardized factor loadings and intercepts for the CFAs are presented in Table 3.

Measurement Invariance

A series of nested multi-group CFAs was used to evaluate three levels of measurement invariance (i.e. configural invariance, metric invariance, and scalar invariance). The first step in testing for measurement invariance is to establish configural invariance. Without first establishing configural invariance, it is not possible to conduct tests of more stringent levels of measurement invariance because the configural invariance model provides a baseline against which subsequent models can be compared (Vandenberg and Lance 2000). Configural invariance indicates that the factor structure is the same for all groups, and is attained if a CFA fits well when the intercepts, factor loadings, and residual variances vary freely across groups, and the factor means are fixed to zero in all groups.

Metric invariance indicates that the factor loadings are the same across all groups. To establish metric invariance, a model is estimated in which factor loadings are constrained to be equal across groups, intercepts and residual variances are free, and factor means are fixed to zero in all groups. The chi-square statistics for the configural invariance model and the metric invariance model can be compared to determine whether the fit for these two models is significantly different; a nonsignificant test indicates that metric invariance is likely to hold.

---

[2] The Kaiser criterion, which states that one should retain factors with eigenvalues greater than 1.0, was not used because it tends to overestimate the true number of factors (Lance et al. 2006; Zwick and Velicer 1982). In addition, the Kaiser criterion should only be used for principal components analysis, not factor analysis (Fabrigar et al. 1999).

Table 3 Intercepts and standardized factor loadings from CFAs, separately for each ethnic group and by sex

Variable Name	White	African American	Hispanic	Male	Female
Intercepts					
Knock Tap	0.62	0.57	0.59	0.56	0.60
CPT	0.94	0.93	0.93	0.93	0.94
Bird/Dragon	3.11	2.83	3.17	2.88	3.07
Gift Wrap	0.70	0.52	0.50	0.51	0.62
Waiting for Bow	1.80	1.59	1.65	1.66	1.66
Yarn Tangle	3.80	3.33	2.71	3.14	3.45
Rabbit Turtle	4.07	3.79	3.84	3.93	3.83
Teacher CBQ	4.49	4.40	4.62	4.27	4.66
Factor Loadings					
Knock Tap	0.38	0.50	0.38	0.49	0.41
CPT	0.44	0.49	0.36	0.36	0.50
Bird/Dragon	0.69	0.51	0.69	0.60	0.55
Gift Wrap	0.68	0.65	0.45	0.58	0.64
Waiting for Bow	0.37	0.30	0.48	0.33	0.40
Yarn Tangle	0.43	0.43	0.31	0.37	0.44
Rabbit Turtle	0.39	0.28	0.22	0.37	0.38
Teacher CBQ	0.53	0.41	0.51	0.42	0.46

Finally, scalar invariance indicates that the intercepts are the same across all groups. For the scalar invariance model, intercepts and factors loadings are constrained to be equal across groups, the residual variances are free, and the factor means are set to zero in one group and free in the others. This was compared using the chi-square difference test of model fit to the metric invariance model. If this test is statistically significant at $\alpha = .05$, full scalar invariance is not achieved. However, partial scalar invariance can still be attained by using modification indices to free individual parameters until the chi-square difference test for model fit is no longer significant; without at least partial scalar invariance, it is not possible to compare group means.

Cross-sex Invariance

The multi-group model testing configural invariance, $\chi^2(40, n = 853) = 72.33$, $p<.01$ $RMSEA = .044$ (90% $CI = .027–.059$), $SRMR = .03$, fit the data well, as did the metric invariance model, $\chi^2(47, n = 853) = 78.25$, $p<.01$, $RMSEA = .039$ (90% $CI = .023–.055$), $SRMR = .062$. Furthermore, the chi-square difference test between the configural invariance and metric invariance model was nonsignificant, $\chi^2(7) = 5.82$, $p = .56$, suggesting that metric invariance between the sexes was attained. Next, the model for scalar invariance was run. This model had acceptable

Table 4 Intercepts and unstandardized factor loadings from multi-group CFA

Variable	Male	Female	European American	African American	Hispanic
Intercepts					
Knock Tap	0.61			0.62	
CPT	0.94			0.94	
Bird/Dragon	3.09			3.12	
Gift Wrap	0.61		0.70[a]	0.62[a]	0.50[a]
Waiting for Bow	1.69		1.80[a]	1.65	1.65
Yarn Tangle	3.41		3.66	3.66	2.71[a]
Rabbit Turtle	4.00[a]	3.83	4.07[a]	3.85	3.85
Teacher CBQ	4.41[a]	4.67		4.56	
Factor Loadings					
Knock Tap	1.00			1.00	
CPT	0.13			0.13	
Bird/Dragon	4.16			4.46	
Gift Wrap	1.64			1.60	
Waiting for Bow	1.03			0.91	
Yarn Tangle	4.09			3.99	
Rabbit Turtle	1.36			1.04	
Teacher CBQ	2.55			2.81	

One value under *Intercepts* or *Factor Loadings* columns means invariant constraints (i.e., the same intercept/factor loading held across the three ethnic groups and across the sexes). Multiple values indicate that the constraint was relaxed.

[a] This intercept value differed statistically from the intercept(s) for the other group(s).

fit, $\chi^2(54, n = 853) = 119.70$, $p<.01$, $RMSEA = .053$ (90% $CI = .041–.066$), $SRMR = .068$, but the chi-square difference test comparing the metric invariance model with the scalar invariance model was statistically significant, $\chi^2(7) = 41.45$, $p<.01$, indicating that full scalar invariance was not met. Allowing the intercepts for *rabbit turtle* and *teacher CBQ* to vary across the sexes as indicated by the modification indices produced a model with good fit, $\chi^2(52, n = 853) = 83.80$, $p<.01$, $RMSEA = .038$ (90% $CI = .022–.052$), $SRMR = .060$, and resulted in a nonsignificant chi-square test for the difference in model fit relative to the metric invariance model, $\chi^2(5) = 5.56$, $p = .35$. The unstandardized factor loadings and intercepts for model establishing partial scalar invariance are displayed in the left part of Table 4.

Cross-ethnic Invariance

The configural invariance model fit the data well, $\chi^2(60, n = 853) = 86.50$, $p<.05$, $RMSEA = .039$ (90% $CI = .018–.057$), $SRMR = .045$, as did the metric invariance

model, $\chi^2(74, n = 853) = 86.50$, $p<.01$, $RMSEA = .041$ (90% $CI = .024–.057$), $SRMR = .070$. The chi-square difference test between these two models was nonsignificant, $\chi^2(14) = 23.59$, $p = .05$, indicating that metric invariance exists between the three ethnic groups. The scalar invariance model fit the data acceptably according to the $RMSEA$, but poorly according to the $SRMR$, $\chi^2(88, n = 853) = 214.08$, $p<.01$, $RMSEA = .071$ (90% $CI = .059–.083$), $SRMR = .092$, and the chi-square difference test for model fit between the metric and scalar invariance models was significant, $\chi^2(14) = 103.98$, $p<.01$. Based on modification indices, the intercepts for waiting for bow and rabbit-turtle in the European-American group were allowed to differ from those for the African American and Hispanic groups, the intercept for yarn tangle in the Hispanic group were relaxed to be different from the African-American and European American groups, and the intercept for gift wrap was allowed to differ in all three groups. In this model, the chi-square difference test between the partial scalar and the metric invariance models was nonsignificant, $\chi^2(9) = 12.12$, $p = .21$. The model for partial scalar invariance fit the data well, $\chi^2(83, n = 853) = 122.21$, $p<.01$, $RMSEA = .041$ (90% $CI = .024–.056$), $SRMR = .069^3$. The unstandardized factor loadings and intercepts for the model establishing partial scalar invariance are displayed in the right part of Table 4.

Because nearly all the European American children were located at the Florida site, and nearly all the Hispanic children were located at the Texas site, site differences could be confounded with ethnic differences. Separate measurement invariance analyses were conducted to examine the measurement equivalence across African Americans at the two sites, and the two major ethnic groups within each site: African Americans and Hispanics in Texas, and African Americans and European Americans in Florida.

The configural invariance model for the African American group across site had acceptable fit, $\chi^2(40, n = 422) = 65.60$, $p<.01$, $RMSEA = .055$ (90% $CI = .029–.078$), $SRMR = .057$, and the metric invariance model had good fit, $\chi^2(47, n = 422) = 71.17$, $p<.05$, $RMSEA = .049$ (90% $CI = .023–.072$), $SRMR = .072$. The chi-square test comparing these two models was nonsignificant, $\chi^2(7) = 5.56$, $p = .59$, indicating that metric invariance held across these two groups. The scalar invariance model had acceptable fit based on the $RMSEA$, but relatively poor fit based on the $SRMR$, $\chi^2(54, n = 422) = 94.92$, $p<.01$, $RMSEA = .060$ (90% $CI = .039–.080$), $SRMR = .090$, and the chi-square difference test comparing the metric invariance model with the scalar invariance model was significant, $\chi^2(7) = 23.75$, $p<.05$, indicating that full scalar invariance was not attained. Releasing the constraints on the intercepts for bird and dragon and the CPT reduced the chi-square difference test for partial scalar invariance to

nonsignificance, $\chi^2(5) = 5.22$, $p = .40$, and produced a model with good fit, $\chi^2(53, n=422) = 76.39$, $p<.05$, RMSEA = .047 (90% CI = .021–.069), SRMR = .072.

The configural invariance model for the Texas site fit well, $\chi^2(40, n = 405) = 55.83$, $p<.05$, RMSEA = .044 (90% CI = .002–.069), SRMR = .051, as did the metric invariance model, $\chi^2(47, n = 405) = 64.94$, $p<.05$, RMSEA = .043 (90% CI = .008–.067), SRMR = .063. The chi-square difference test for these models was nonsignificant, $\chi^2(7) = 9.12$, $p = .24$, indicating that metric invariance held for these two groups. The full scalar invariance model fit the data adequately, $\chi^2(54, n = 405) = 96.76$, $p<.01$, RMSEA = .062 (90% CI = .041–.082), SRMR = .071. The chi-square difference test between the metric invariance model and the scalar invariance model was significant, $\chi^2(7) = 31.82$, $p<.01$, indicating that the requirements for full scalar invariance were not met. Relaxing the constraint on the intercept for yarn tangle produced a model with good fit, $\chi^2(53, n = 405) = 73.95$, $p<.05$, RMSEA = .04 (90% CI = .014–.066), SRMR = .061, and produced a nonsignificant chi-square difference test, $\chi^2(6) = 9.01$, $p = .17$.

Similarly, the configural model for the Florida site also fit well, $\chi^2(40, n = 421) = 44.61$, $p = .28$, RMSEA = .023 (90% CI = .000–.044), SRMR = .049. The RMSEA statistic indicated the metric invariance model also fit well, although the SRMR statistic for this model was above .08, indicating a relatively poor fit, $\chi^2(47, n = 421) = 62.62$, $p = .06$, RMSEA = .04 (90% CI = .000–.064), SRMR = .085. The chi-square difference test for this model was significant, $\chi^2(7) = 18.01$, $p<.05$, indicating that full metric invariance was not met between ethnic groups in Florida. The constraint on the correlation between yarn tangle and waiting for bow was released, producing a model with a slightly improved SRMR statistic, $\chi^2(40, n = 421) = 56.91$, $p = .13$, RMSEA = .034 (90% CI = .000–.059), SRMR = .080. This model fit as well as the configural invariance model, $\chi^2(6) = 12.30$, $p = .06$, establishing partial metric invariance. The constraints for scalar invariance were added to the partial metric invariance model, producing a model with a good RMSEA value but a high SRMR value, $\chi^2(53, n = 421) = 74.11$, $p<.05$, RMSEA = .043 (90% CI = .014–.066), SRMR = .094. The chi-square difference test comparing this model the partial metric invariance model was significant, $\chi^2(7) = 17.20$, $p<.05$. Releasing the constraint on the intercept for rabbit turtle did little to improve the SRMR statistic, $\chi^2(52, n = 421) = 66.79$, $p = .08$, RMSEA = .037 (90% CI = .000–.060), SRMR = .089, but produced a nonsignificant chi-square difference test, $\chi^2(6) = 9.87$, $p = .13$.

The results of these analyses indicated that there were relatively few differences between the groups when analyzed separately by site, and that full scalar invariance was not established across sites for the African American group. Thus, differences between the sites may partially account for the differences in intercepts between groups in the partial scalar invariance model using data from both sites.[3]

Discussion

The purpose of this study was to examine the factor structure and measurement invariance of EC across three ethnic groups (European Americans, African Americans, and Hispanics) and across sex. In a factor analysis of teachers' reports of the inhibitory and attentional focusing scales of the CBQ and seven behavioral measures thought to index EC, we found that a single factor solution best accounted for the pattern of correlations among the variables. Tests of configural and metric invariance indicated that the factor structure and the pattern of loadings were not significantly different across boy and girls or across the three ethnic groups studied. That is, a single factor model fit the data for each group equally well, and the factor loadings were not different among the groups.

Of our hypothesized indicators of EC, bird and dragon and knock tap are most indicative of inhibitory control, yarn tangle (persistence on a boring task) probably taps activational and perhaps attentional control, the rabbit and turtle task reflects motor, inhibitory, and activational control, the CPT is a measure of attentional control, and gift wrap and waiting for bow likely tap impulsivity or sensitivity to reward in addition to inhibitory (and perhaps attentional) control. Thus, our battery of indicators included a diverse set of skills encompassed by EC at preschool age, including gross motor control, inhibitory and activational control, attention focusing, and the ability to delay gratification. Each of these subcomponents of EC involves a somewhat different skill set, but they are typically viewed as indicators of a more general construct, and are commonly combined to form a single composite measure of EC. Our findings support the conclusion that all of these constructs tap the common latent construct of EC. Even though there was some evidence for two factors in some auxiliary analyses (see Footnote 2), the two factors were always highly correlated. This is an important finding

---

[3] Measurement invariance models in which the Hispanic group consisted of only of English speaking students or only Spanish-speaking students were compared to the models that included the entire sample. In these alternate models, metric, configural, and partial scalar invariance were established in both language groups. In the full sample, 5 intercepts needed to be released for partial scalar invariance. In the English speaking group, 4 needed to be released and all were the same as in the full sample. In the Spanish-speaking group, 5 needed to be released: 3 were the same as the full sample, and two were different. This finding suggests that language differences cannot fully explain the differences in intercepts among the ethnic groups in our sample.

because it was unclear if some of the tasks might measure impulsivity or behavioral inhibition (i.e., reactive control) more than EC. Although it is quite possible that some of the indices do tap impulsivity to some degree, they also appear to measure EC in a manner that is consistent with other indicators of EC. However, to our knowledge there is no evidence that any specific aspect of EC has better predictive validity than the more general factor.

In addition to full configural and metric invariance, we established partial scalar invariance across sex. Only two intercepts differed for males and females: the intercept for rabbit turtle was larger in males, whereas the intercept for the teacher-rated EC on the CBQ was larger in females. The difference in intercept for the CBQ was expected, as gender stereotypes likely influence teachers' perceptions of EC above and beyond actual differences between boys' and girls' behavior. Given the relatively small number of differences between these two groups, we have considerable confidence that means can be compared across gender, especially if questionnaire measures are not used as indicators of EC.

Configural, metric, and partial scalar invariance were also established across ethnic groups. In the multi-group model comparing European Americans, African Americans, and Hispanics, several differences in intercepts emerged. The intercepts for waiting for bow and rabbit turtle were larger in European American group relative to the other two groups. For gift wrap, the intercept was largest in the European American group, and smallest in the Hispanic group, with the African American group falling in the middle. Finally, the intercept for yarn tangle was smallest in the Hispanic group relative to the other two groups. The intercepts for the other four indicators did not differ across groups.

Unfortunately, we cannot offer any theoretical explanation for the differences in intercepts. European Americans and Hispanics were most dissimilar, with four intercepts that differed; The African American group had two intercepts that differed from this Hispanic group, and one that differed from the European American group. The differences in intercepts may have been due to cultural differences between the ethnic groups, or to cultural or individual differences between participants in Texas and Florida that are unrelated to ethnicity. The African American group was split between Texas and Florida, but nearly all of the Hispanic participants were in Texas, and nearly all of the European American participants were in Florida. Analyses of invariance conducted within each site revealed few differences in intercepts, and suggest that differences between groups in intercepts may be partially due to site differences rather than ethnic differences. This conclusion is supported by the larger number of differences between European Americans and Hispanics, groups that were primarily located at two different sites, than between those groups and African Americans, who were split between the sites.

It is also possible that the ethnic groups differed in income or other variables, including familial or school socialization, that we did not measure or that the tasks had differential appeal to the children in various groups. Differences in maternal education or socioeconomic status more generally likely also account for some of the differences in the intercepts; the Hispanic group was lowest in maternal education and had the lowest scores on several variables. Thus, we were not able determine to what degree socioeconomic, cultural, and other factors were related to differences in intercepts across ethnic groups.

To our knowledge, this is the first study to investigate the factor structure of EC in a large, high-risk, low-income sample with significant numbers of ethnic minorities. We have established configural and metric invariance for EC across the sexes and across three ethnic groups using a diverse array of behavioral indicators and teacher reports. This indicates that the construct of EC behaves in a similar way across groups, and that a wide array of tasks index a single latent EC construct. Partial scalar invariance was also established, indicating that comparisons of mean levels of EC are valid between girls and boys, and, although there were several differences among the intercepts, comparisons of mean levels across ethnicity have some validity as well. However, if investigators use only one specific measure of EC rather than an aggregate, it is possible that the intercept may not be equivalent across groups; in this case, comparisons of mean levels may not be meaningful. A goal for future research is to examine whether the predictive validity of measures of EC is equivalent across groups. In addition, it would be useful to examine the measurement invariance of EC across different levels of SES and in a range of cultures, including groups outside of the United States.

References

Aikens, N. L., Coleman, C. P., & Barbarin, O. A. (2008). Ethnic differences in the effects of parental depression on preschool children's socioemotional functioning. *Social Development, 17*, 137–160.

Blair, C., Granger, D., & Razza, R. P. (2005). Cortisol reactivity is positively related to executive function in preschool children attending Head Start. *Child Development, 76*, 554–554.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & S. J. Long (Eds.), *Testing structural equation models*. Newbery Park: Sage.

Byrne, B. M. (1994). Testing for the factorial validity, replication, and invariance of a measuring instrument: A paradigmatic application based on the Maslach Burnout Inventory. *Multivariate Behavioral Research, 29*, 289–311.

Capaldi, D. M., & Rothbart, M. K. (1992). Development and validation of an early adolescent temperament measure. *The Journal of Early Adolescence, 12*, 153–173.

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

Carver, C. S. (2005). Impulse and constraint: Perspectives from personality psychology, convergence with theory in other areas, and potential for integration. *Personality and Social Psychology Review, 9*, 312–333.

Derryberry, D., & Rothbart, M. K. (1997). Reactive and effortful processes in the organization of temperament. *Development & Psychopathology, 9*, 633–652.

Eisenberg, N., & Morris, A. S. (2002). Children's emotion-related regulation. In R. V. Kail (Ed.), *Advances in child development and behavior* (Vol. 30, pp. 189–229). San Diego: Academic.

Eisenberg, N., Fabes, R. A., Karbon, M., Murphy, B. C., Wosinski, M., Polazzi, L., et al. (1996). The relations of children's dispositional prosocial behavior to emotionality, regulation, and social functioning. *Child Development, 67*, 974–992.

Eisenberg, N., Fabes, R. A., Shepard, S. A., Murphy, B. C., Jones, J., & Guthrie, I. K. (1998). Contemporaneous and longitudinal prediction of children's sympathy from dispositional regulation and emotionality. *Developmental Psychology, 34*, 910–924.

Eisenberg, N., Pidada, S., & Liew, J. (2001). The relations of regulation and negative emotionality to Indonesian children's social functioning. *Child Development, 72*, 1747–1763.

Eisenberg, N., Spinrad, T. L., Fabes, R. A., Reiser, M., Cumberland, A., Shepard, S. A., et al. (2004). The relations of effortful control and impulsivity to children's resiliency and adjustment. *Child Development, 75*, 25–46.

Eisenberg, N., Sadovsky, A., & Spinrad, T. L. (2005a). Associations among emotion-related regulation, language skills, emotion knowledge, and academic outcomes. *New Directions in Child and Adolescent Development, 109*, 109–118.

Eisenberg, N., Sadovsky, A., Spinrad, T. L., Fabes, R. A., Losoya, S. H., Valiente, C., et al. (2005b). The relations of problem behavior status to children's negative emotionality, effortful control, and impulsivity: Concurrent relations and prediction of change. *Developmental Psychology, 41*, 193–211.

Eisenberg, N., Ma, Y., Chang, L., Zhou, Q., West, S. G., & Aiken, L. (2007a). Relations of effortful control, reactive undercontrol, and anger to Chinese children's adjustment. *Development and Psychopathology, 19*, 385–409.

Eisenberg, N., Michalik, N., Spinrad, T. L., Hofer, C., Kupfer, A., Valiente, C., et al. (2007b). The relations of effortful control and impulsivity to children's sympathy: A longitudinal study. *Cognitive Development, 22*, 544–567.

Else-Quest, N. M., Hyde, J. S., Goldsmith, H. H., & Van Hulle, C. A. (2006). Gender differences in temperament: A meta-analysis. *Psychological Bulletin, 132*, 33–72.

Evans, G. W., & English, K. (2002). The environment of poverty: multiple stressor exposure, psychophysiological stress, and socioemotional adjustment. *Child Development, 73*, 1238–1248.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272–299.

Fantuzzo, J., Grim, S., Mordell, M., McDermott, P., Miller, L., & Coolahan, K. (2001). A multivariate analysis of the revised Conners' teacher rating scale with low-income, urban preschool children. *Journal of Abnormal Child Psychology, 29*, 141–152.

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

Gilliom, M., Shaw, D. S., Beck, J. E., Schonberg, M. A., & Lukon, J. L. (2002). Anger regulation in disadvantaged preschool boys: Strategies, antecedents, and the development of self-control. *Developmental Psychology, 38*, 222–235.

Goldsmith, H. H., Reilly, J., Lemery, K. S., Longley, S., & Prescott, A. (1993). *Preschool Laboratory Temperament Assessment Battery (*PS Lab-TAB; Version 1.0). Technical Report, Department of Psychology, University of Wisconsin-Madison.

Howse, R. B., Lange, G., Farran, D. C., & Boyles, C. D. (2003). Motivation and self-regulation as predictors of achievement in economically disadvantaged young children. *Journal of Experimental Education, 71*, 151–174.

Ispa, J. M., Fine, M. A., Halgunseth, L. C., Harper, S., Robinson, J., Boyce, L., et al. (2004). Maternal intrusiveness, maternal warmth, and mother-toddler relationship outcomes: Variations across low-income ethnic and acculturation groups. *Child Development, 75*, 1613–1631.

Kagan, J., & Fox, N. (2006). Biology, culture, and temperamental biases. In W. Damon, R. L. Lerner (Series Eds.), & N. Eisenberg (Vol. Ed.), *Handbook of child psychology*: *Vol. 3: Social, emotional, and personality development* (6th ed., pp. 167–225). New York: Wiley.

Kelloway, E. K. (1998). *Using LISREL for structural equation modeling: A researcher's guide*. Thousand Oaks: Sage.

Kim, S., Brody, G. H., & Murry, V. M. (2003). Factor structure of the early adolescent temperament questionnaire and measurement invariance across gender. *The Journal of Early Adolescence, 23*, 268–294.

Kindlon, D., Mezzacappa, E., & Earls, F. (1995). Psychometric properties of impulsivity measures: Temporal stability, validity, and factor structure. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 36*, 645–661.

Kline, R. B. (1998). *Principles and practices of structural equation modeling*. New York: Guilford.

Kochanksa, G., & Knaack, A. (2003). Effortful control as a personality characteristic of young children: Antecedents, correlates, and consequences. *Journal of Personality, 71*, 1087–1112.

Kochanska, G., Murray, K. T., & Harlan, E. T. (2000). Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. *Developmental Psychology, 36*, 220–232.

Kochanska, G., Murray, K., Jacques, T. Y., Koenig, A. L., & Vandegeest, K. A. (1996). Inhibitory control in young children and its role in emerging internalization. *Child Development, 67*, 490–507.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods, 9*, 202–220.

Lengua, L. J. (2003). Associations among emotionality, self-regulation, adjustment problems, and positive adjustment in middle childhood. *Journal of Applied Developmental Psychology, 24*, 595–618.

Li-Grining, C. P. (2007). Effortful control among low-income preschoolers in three cities: Stability, change, and individual differences. *Developmental Psychology, 43*, 208–221.

Loukas, A., & Roalson, L. A. (2006). Family environment, effortful control, and adjustment among European American and Latino early adolescents. *Journal of Early Adolescence, 26*, 432–455.

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

Martel, M. M., Nigg, J. T., Wong, M. M., Fitzgerald, H. E., Jester, J. M., Puttler, L. I., et al. (2007). Childhood and adolescent resiliency, regulation, and executive functioning in relation to adolescent problems and competence in a high-risk sample. *Development and Psychopathology, 19*, 541–563.

McCabe, L. A., Rebello-Britto, P., Hernandez, M., & Brooks-Gunn, J. (2004). Games children play: Observing young children's self-regulation across laboratory, home, and school settings. In R. DelCarmen-Wiggins & A. Carter (Eds.), *Handbook of infant,*

toddler, and preschool mention health assessment (pp. 491–521). New York: Oxford University Press.

Mendez, J. L., Fantuzzo, J., & Cicchetti, D. (2002). Profiles of social competence among low-income African American preschool children. *Child Development, 73*, 1085–1100.

Mezzacappa, E. (2004). Alerting, orienting, and executive attention: developmental properties and sociodemographic correlates in an epidemiological sample of young, urban children. *Child Development, 75*, 1373–1386.

Miller, A. L., Fine, S. E., Gouley, K. K., Seifer, R., Dickstein, S., & Shields, A. (2006). Showing and telling about emotions: interrelations between facets of emotional competence and associations with classroom adjustment in Head Start preschoolers. *Cognition & Emotion, 20*, 1170–1170.

Muris, P., de Jong, P. J., & Engelen, S. (2004). Relationships between neuroticism, attentional control, and anxiety disorders symptoms in non-clinical children. *Personality and Individual Differences, 37*, 789–797.

Murray, K. T., & Kochanska, G. (2002). Effortful control: factor structure and relation to externalizing and internalizing behaviors. *Journal of Abnormal Child Psychology, 30*, 503–514.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide* (5th ed.). Los Angeles: Muthén & Muthén.

Nigg, J. T. (2000). On inhibition/disinhibition in developmental psychopathology: Views from cognitive and personality psychology and a working inhibition taxonomy. *Psychological Bulletin, 126*, 230–246.

Noble, K. G., Norman, M. F., & Farah, M. J. (2005). Neurocognitive correlates of socioeconomic status in kindergarten children. *Developmental Science, 8*, 74–87.

Oldehinkel, A. J., Hartman, C. A., Ferdinand, R. F., Verhulst, F. C., & Ormel, J. (2007). Effortful control as a modifier of the association between negative emotionality and adolescents' mental health problems. *Development and Psychopathology, 19*, 523–539.

Olson, S. L., Bates, J. E., & Bayles, K. (1990). Early antecedents of childhood impulsivity: The role of parent-child interaction, cognitive competence, and temperament. *Journal of Abnormal Child Psychology, 18*, 317–334.

Olson, S. L., Sameroff, A. J., Kerr, D. C. R., Lopez, N. L., & Wellman, H. M. (2005). Developmental foundations of externalizing problems in young children: The role of effortful control. *Development and Psychopathology, 17*, 25–45.

Olson, S. L., Schilling, E. M., & Bates, J. E. (1999). Measurement of impulsivity: Construct coherence, longitudinal stability, and relationship with externalizing problems in middle childhood and adolescence. *Journal of Abnormal Child Psychology, 27*, 151–165.

Perner, J., & Lang, B. (2000). Theory of mind and executive function: Is there a developmental relationship? In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience* (2nd ed., pp. 150–181). New York: Oxford University Press.

Pickering, A. D., & Gray, J. A. (1999). The neuroscience of personality. In O. P. John & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 277–299). New York: Guilford.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163.

Raver, C. C. (2004). Placing emotional self-regulation in sociocultural and socioeconomic contexts. *Child Development, 75*, 346–353.

Reed, M. A., Pien, D. L., & Rothbart, M. K. (1984). Inhibitory self-control in preschool children. *Merrill-Palmer Quarterly, 30*(2), 131–147.

Riggs, N. R., Greenberg, M. T., Kusche, C. A., & Pentz, M. A. (2006). The mediational role of neurocognition in the behavioral outcomes of a social-emotional prevention program in elementary school

students: Effects of the PATHS curriculum. *Prevention Science, 7*, 91–102.

Rosvold, H. E., Mirsky, A. F., Sarason, I, Bransome, E. D., Jr., & Beck, L. H. (1956). A continuous performance test of brain damage. *Journal of Consulting Psychology, 20*(5), 343–350.

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

Rothbart, M. K., & Bates, J. E. (2006). Temperament. In W. Damon, R. L. Lerner (Series Eds.), & N. Eisenberg (Vol. Ed.), *Handbook of child psychology*: *Vol. 3: Social, emotional, and personality development* (6th ed., pp. 99–166). New York: Wiley.

Rydell, A.-M., Berlin, L., & Bohlin, G. (2003). Emotionality, emotion regulation, and adaptation among 5- to 8-year-old children. *Emotion, 3*, 30–47.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods, 7*, 147–177.

Silk, J. S., Steinberg, L., & Morris, A. S. (2003). Adolescents' emotion regulation in daily life: Links to depressive symptoms and problem behavior. *Child Development, 74*, 1869–1880.

Spinrad, T. L., Eisenberg, N., Cumberland, A., Fabes, R. A., Valiente, C., Shepard, S., et al. (2006). Relation of emotion-related regulation to children's social competence: A longitudinal study. *Emotion, 6*, 498–510.

Spinrad, T. L., Eisenberg, N., Gaertner, B., Popp, T., Smith, C. L., Kupfer, A., et al. (2007). Relations of maternal socialization and toddlers' effortful control to children's adjustment and social competence. *Developmental Psychology, 43*, 1170–1186.

Tabachnik, B. G., & Fidell, L. S. (2006). *Using multivariate statistics* (5th ed.): Pearson/Allyn & Bacon.

Valiente, C., Eisenberg, N., Smith, C. L., Reiser, M., Fabes, R. A., Losoya, S., et al. (2003). The relations of effortful control and reactive control to children's externalizing problems: A longitudinal assessment. *Journal of Personality, 71*, 1171–1196.

Valiente, C., Lemery-Chalfant, K., Swanson, J., & Reiser, M. (2008). Prediction of children's academic competence from their effortful control, relationships, and classroom participation. *Journal of Educational Psychology, 100*, 67–77.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–69.

White, J. L., Moffitt, T. E., Caspi, A., Bartusch, D. J., Needles, D. J., & Stouthamer-Loeber, M. (1994). Measuring impulsivity and examining its relationship to delinquency. *Journal of Abnormal Psychology, 103*, 192–205.

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods, 8*, 16–37.

Windle, M. (1992). Revised Dimensions of Temperament Survey (DOTS-R): simultaneous group confirmatory factor analysis for adolescent gender groups. *Psychological Assessment, 4*, 228–234.

Windle, M., & Lerner, R. M. (1986). Reassessing the dimensions of temperamental individuality across the life span: The revised dimensions of temperament survey (DOTS-R). *Journal of Adolescent Research, 1*, 213–229.

Zhou, Q., Wang, Y., Eisenberg, N., Wolchik, S., Tein, J.-W., & Deng, X. (2008). Relations of parenting and temperament to Chinese children's experience of negative life events, coping efficacy, and externalizing problems. *Child Development, 79*, 493–513.

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research, 17*, 253–269.