The following Supplementary Materials are relevant for the method and results described in the main text.

**Method**

**Procedures**

Participants were recruited through several means, including flyers posted at various locations and word of mouth. Children were compensated for participation with a small toy and/or book, along with a certificate describing them as a “junior scientist.” Parents were compensated with a $10 gift card and a DVD copy of their child completing the tasks. A phone interview was conducted prior to the study to determine eligibility and child language background.

After signing informed consent, the parent was taken to a separate room to complete three surveys (they could see the child through a monitor). In the experiment room on campus, a bilingual RA built rapport with the child and gained assent.

**Measures**

***Language Background***

**Expressive Language Skills.** Wordless picture books have been used frequently in prior studies to elicit narratives from young children ranging in age from 3 to 10 years (Cain, 2003; de Villiers, 2004; Symons et al., 2005; Bedore et al., 2010; Heilmann et al., 2008; Justice et al., 2010) and demonstrate high ecological validity (Bedore et al., 2010; Westby et al., 1989).

The procedure for the story-telling task followed the Narrative Elicitation Protocol outlined by Justice and colleagues (2010). This elicitation protocol is part of a larger Narrative Assessment Protocol described by Justice et al., but which was not used in the current study. To begin, the RA told the child that the RA would go through the pictures in the book once with the child, and then it would be the child’s turn to create a make-believe story about the book. The RA then went through the book once with the child, pointing out the pictures, and spending eight to ten seconds on each page. The RA did not make any comments about the content of any of the pictures. If the participant made a comment, the RA reminded the child that they were only looking at the pictures at that time. After going through the book one time, the RA handed the child the book and asked him/her to tell a make-believe story using the pictures in the book. The child was told to make the story as long and interesting as possible (Justice et al., 2010). While the child told his/her story, the RA only provided positive feedback (“Good job!” or “I like your story!”), repeated exactly what the child had just said, or prompted the child to continue if the child appeared stuck on a page. When the child reached the end, the RA asked if the child had anything else to add. For bilingual participants, the RA repeated the task, including all instructions and feedback, in the opposite language.

For total number of words, any and all intelligible words the child uttered were part of this study’s code category, including grammatical errors, repeated words and restarts, but excluding filler words like “um” or “uh” or words/phrases that were clearly directed to the self or to the experimenter and were not part of the story (ex: child says “What is that?” to self or experimenter about a picture in the book). NDW only includes unique words the child uttered across the entire transcript, which reflect children’s language growth (Bedore et al., 2010). For instance, once a child said “frog” in his/her story, the word “frog” would not count again toward the NDW total. Two RAs coded the same five transcripts and compared results and discussed to resolve any discrepancies.

**Language Experience and Exposure.** The survey also asked parents about the child’s exposure to the two languages. Each statement required a response ranging on a 1 (always English) to 5 (always Spanish) scale, reflecting actual language use by the child in the everyday environment. Seven statements reflected language output for the child and seven reflected language input. Scores on the first fourteen statements (reflecting balance between the two languages on input and output) were averaged together to get one score for children’s language use/balance. Based on recommendations from prior research that it is important to focus on bilingual children’s actual language use (Morton & Harper, 2007), a more stringent language use score was calculated based on only six of the seven language output items. So, because this scale ranged from 1 to 5, a score in the range of 2.5 to 3.5 indicated almost complete balance between the two languages in terms of use. A score less than 2.5 indicated greater English language use by the child, and a score above 3.5 indicated greater Spanish language use.

**Determining Child Language Group.** The criteria for language group were created based on prior research. First, when thinking about *how* bilingual an individual is, researchers have noted the importance of considering length of exposure, proficiency or ability in each language, and actual use of the two languages (Baker 2006; Bialystok, 2001). Criterion 1 taps into length of exposure, criteria 2, 4, 5, and 6 tap into language proficiency/ability, and criterion 3 taps into actual language usage. Second, the minimum requirement of three years of exposure has been used as a criterion in other studies (Guitérrez-Clellen & Kreiter, 2003), and is based on findings from Cummins (1984) that three to four years of exposure to a language is required “to achieve basic social language competencies” (Guitérrez-Clellen & Kreiter, 2003, p. 271). Third, the computation of balance between the two languages by averaging scores across the parent-reported statements of language use is based on prior work by Bialystok (Bialystok, 2010; Bialystok & Feng, 2009).

**Results**

***Descriptive***

We first examined bivariate relations between EF and demographic variables, including child gender, age, ethnicity, immigrant status, total children in the family, family income, mother/father education, family structure, and PPVT score. Gender, ethnicity, immigrant status, family income, mother/father education, and PPVT scores were not related to any EF variable. Child age was related to DCCS errors (r = -.26, p = .02), GNG commission errors (r = -.25, p = .03), and parent reported BRIEF BRI—older children made fewer errors on both tasks and had fewer behavioral/EF problems. Number of children in the family was positively related to DCCS errors (r = .25, p = .046) and negatively related to Simon effect score (r = -.33, p = .008), indicating that children from larger households had more errors on the DCCS, but better interference control.

We also examined the inter-correlations between all EF measures. GNG commission errors were positively related to parent BRIEF reports—children who performed worse on the GNG (poorer inhibitory control) were rated by parents as having more EF deficits. RT on Simon congruent trials was correlated with RT on incongruent trials, meaning that children with slower RT on one trial type were likely to have slower RTs on the other trial type. Unexpectedly, the Simon Effect was negatively related to parent report of Metacognition Index (MI) and Global Executive Composite (GEC) but not Behavior Regulation Index (BRI). Children with a higher Simon Effect (i.e., less interference control) were rated by their parents as having fewer EF deficits. The DCCS was positively related to parent report of BRI but not MI and the GEC, suggesting that poorer cognitive flexibility was related to parents reporting more behavioral but not metacognition issues for the child. The BRIEF indexes (BRI and MI) and composite (GEC) were highly correlated with one another, as to be expected.

Other relations were marginally significant. GNG commission errors were negatively related to HTKS score and positively related to DCCS errors (r = -.23, p = .053)—more errors on the GNG (poorer inhibitory control) were associated with worse performance on both the HTKS (also poorer inhibitory control) and DCCS (poorer cognitive flexibility). RT on congruent trials of the Simon task was positively related to DCCS errors (r = .20, p = .075)—poorer interference control (Simon) was associated with poorer cognitive flexibility (DCCS; larger numbers indicate more errors). Finally, RT on incongruent trials of the Simon task was negatively related with the BRIEF MI (r = -.22, p = .054), indicating that poorer interference control was associated with less parent-reported metacognitive issues.

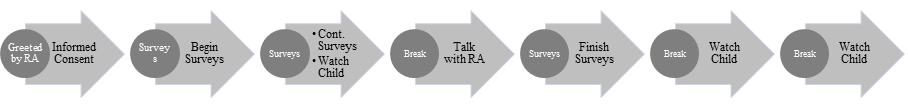
***EF Development and Theory***

Several of the EF measures were related to one another in expected ways. Children with less ability to inhibit a prepotent response (i.e., suppress a dominant response when cued to do so) had more difficulty with other forms of impulse control (i.e., inhibition of a dominant response and replacing it with a less-dominant one), as well as poorer cognitive flexibility. Longer reaction time (RT) on congruent trials of the Simon task, potentially an indicator of poorer interference control (Lu & Proctor, 2001), was related to poorer cognitive flexibility, as well as less accurate problem-solving. There was some evidence for the validity of the parent rating of EF, as it correlated positively with commission errors on the GNG: more deficits in EF as measured by parent report of behavior at home were related to more deficits in inhibitory control as measured by direct assessment. However, parent report of fewer EF problems was unexpectedly related to poorer interference control. The reason for this last finding is unclear. Moreover, various EF components assessed in the current study were somewhat correlated with one another but not to a high degree, implying that multiple distinct components were being assessed. These findings somewhat support the theory of EF as an umbrella construct with multiple inter-related components (Anderson, 2002; Miyake et al., 2000), but it is clear that more theoretical and instrument-validation work is needed (McCoy, 2019; Wallisch et al., 2018).

**Figure S1**

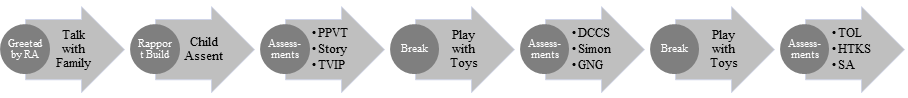
Order of Experimental Procedures for Parent and Child

*Procedures for Parent*



(5 min) (5 min) (20-30 min) (5 min) (30 min) (5 min) (10 min)

*Procedures for Child*



Note: Surveys = Language Background Questionnaire, Behavior Rating Inventory of Executive Function, Child Behavior Checklist; PPVT = Peabody Picture Vocabulary Test; Story = Story telling task (one for monolingual, two for bilingual); TVIP = Test de Vocabulario en Imágenes Peabody; DCCS = Dimensional Change Card Sort; Simon = Simon task; GNG = go/no-go task; HTKS = Heads-toes-knees-shoulders task; SA = selective attention task (not analyzed in current dissertation)

**Table S1**

Description of all measures of EF, Behavior, and Language

|  |  |  |
| --- | --- | --- |
| Assessment | Abbr. | Purpose |
| **Language Assessments** |  |  |
| Peabody Picture Vocabulary Test | PPVT | English receptive language skills |
| Test de Vocabulario en Imágenes Peabody | TVIP | Spanish receptive language skills |
| Narrative Elicitation Task |  | Spanish and/or English productive language skills |
| Language Background Questionnaire |  | Background of language experience (parent report) |
| **Executive Function Assessments** |  |  |
| Dimensional Change Card Sort | DCCS | Cognitive flexibility (set shifting) |
| Simon Task |  | Interference control |
| Go/No-Go Task | GNG | Inhibition of a prepotent response |
| Head-Toes-Knees-Shoulders Task | HTKS | Inhibition of a prepotent response |
| Behavior Rating Inventory of Executive Function | BRIEF | Overall executive function (parent report) |
| **Behavior Assessment** |  |  |
| Child Behavior Checklist | CBCL | Behavior problems (parent report) |

**Table S2**

Bivariate correlations between all EF outcome measures

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1. GNG | -- |  |  |  |  |  |  |  |
| 2. HTKS | -.23+ | -- |  |  |  |  |  |  |
| 3. Cong RT | -.08 | -.05 | -- |  |  |  |  |  |
| 4. Incong RT | -.08 | -.05 | .90\*\* | -- |  |  |  |  |
| 5. Simon Effect | -.09 | -.04 | -.05 | .40\*\* | -- |  |  |  |
| 6. DCCS | .19+ | -.19 | .20+ | .18 | .06 | -- |  |  |
| 7. BRIEF BRI | .34\*\* | -.07 | .12 | .06 | -.19 | .23\* | -- |  |
| 8. BRIEF MI | .26\* | -.22+ | -.11 | -.22+ | -.26\* | -.01 | .53\*\* | -- |
| 9. BRIEF GEC | .31\*\* | -.18 | -.02 | -.12 | -.24\* | .15 | .78\*\* | .90\*\* |

*Note.* \*\* *p* < .01, \**p* < .05, +*p* < .10

GNG = go/no-go, HTKS = Head-to-Toes, Cong/Incong RT = Simon congruent/incongruent trials reaction time, DCCS = Dimensional Change Card Sort, BRIEF = Behavior Rating Inventory of Executive Function, BRI = Behavior Regulation Index, MI = Metacognition Index, GEC = Global Executive Composite