Bilingualism: Language and Cognition

cambridge.org/bil

Working memory structure in young Spanish–English bilingual children

Mary Alt[1], DeAnne R. Hunter[2], Roy Levy[2], Sarah Lynn Neiling[1], Kimberly Leon[1], Genesis D. Arizmendi[1], Nelson Cowan[3] and Shelley Gray[2]

[1]Department of Speech, Language, and Hearing Sciences, University of Arizona, Tucson, AZ, USA; [2]College of Health Solutions, Department of Psychology, Arizona State University, Tempe, AZ, USA and [3]Psychological Sciences, University of Missouri-Columbia, Columbia, MO, USA

Research Article

Cite this article: Alt, M., Hunter, D.R., Levy, R., Neiling, S.L., Leon, K., Arizmendi, G.D., Cowan, N. and Gray, S. (2025). Working memory structure in young Spanish–English bilingual children. *Bilingualism: Language and Cognition* **28**, 469–483. https://doi.org/10.1017/S1366728924000580

Received: 08 September 2023
Revised: 12 June 2024
Accepted: 13 June 2024
First published online: 13 December 2024

Keywords:
bilingual; working memory; invariance testing; models; children

Corresponding author:
Mary Alt;
Email: malt@arizona.edu

This article has earned badge for transparent research practices: Open Data Badge. For details see the Data Availability Statement.

Abstract

Working memory encompasses the limited incoming information that can be held in mind for cognitive processing. To date, we have little information on the effects of bilingualism on working memory because, absent evidence, working memory tasks cannot be assumed to measure the same constructs across language groups. To garner evidence regarding the measurement equivalence in Spanish and English, we examined second-grade children with typical development, including 80 bilingual Spanish–English speakers and 167 monolingual English speakers in the United States, using a test battery for which structural equation models have been tested – the *Comprehensive Assessment Battery for Children – Working Memory* (CABC-WM). Results established measurement invariance across groups up to the level of scalar invariance.

Highlights

- First test of invariance for a working memory model between monolingual English and bilingual Spanish/English children.
- Both groups fit the same three-factor model of working memory to the level of scalar invariance.
- Group differences were affected by socioeconomic status and any group differences are difficult to interpret due to wide probability ranges for likely outcomes.

1. Introduction

There is limited information about the structure of working memory in Spanish–English bilingual children. It is important to understand the nature of bilingual working memory structure for both theoretical and practical reasons. Most available research investigates working memory in monolingual speakers, leaving a gap in our understanding of bilingual working memory structure.

We use working memory to hold information in mind to do cognitive tasks (e.g., Cowan, 2022). It is necessary for daily functions (e.g., efficient grocery shopping) and is a strong predictor of academic achievement (e.g., Ahmed et al., 2019; Cowan, 2014; Peng & Kievit, 2020; Simone et al., 2018; Swanson et al., 2021). As such, it is important to understand how children's working memory is structured, so we can support areas of strengths and weaknesses in working memory. Theoretical and statistical accounts for monolinguals have converged into a factor model of working memory structure that includes the following three components: a central executive component, a phonological factor, and a visuospatial factor (Baddeley & Hitch, 1974; Cowan, 1988). In Gray et al. (2017), the visuospatial factor and the focus of attention factors were combined.

In this study, which was part of the larger project titled, 'Profiles of Working Memory and Word Learning for Educational Research (POWWER)', we aimed to better understand working memory structure and performance in Spanish–English bilingual children. We examined whether the working memory structure for bilingual children also fit the three-factor model of working memory derived by Gray et al. (2017) for monolingual English-speaking children the same age (which was also part of the POWWER project). We administered the *Comprehensive Assessment Battery for Children–Working Memory* (CABC-WM, Gray et al., n.d.; summarized by Cabbage et al., 2017) to simultaneous bilingual Spanish–English second graders and, using data from monolingual English-speaking second graders who completed these same tasks, tested for between-group measurement invariance. If established, this permits us to proceed with between-group comparisons of working memory performance by monolingual and bilingual children.

1.1. Structure and performance levels in working memory measurement: The need for invariance testing

Considerable research has focused on bilingual working memory in the last two decades (see Grundy & Timmer, 2017; Gunnerud et al., 2020; Monnier et al., 2022 for meta-analyses). Understanding how working memory performance compares between monolingual and bilingual populations is of significant educational value. However, previous work conducting such comparisons has not taken the critical first step to evaluate whether the *measurement quality* of working memory tasks in these groups is comparable before proceeding to comparisons between groups in terms of the constructs themselves.

Factor analysis is a statistical approach to characterizing the measurement quality of the working memory tasks in terms of the *factor structure*, by which we mean the pattern and strength of dependence of the observed variables capturing the performance on the tasks on underlying working memory constructs represented by latent variables (aka factors). Factor analysis also provides a framework for invariance testing, to characterize the comparability (or lack thereof) of the measurement quality between groups (e.g., monolingual and bilingual).

Conceptually, there are two competing explanations for why we may observe differences between groups in terms of performance on working memory tasks: (1) there are differences in underlying working memory constructs between the groups and (2) there are differences in the measurement quality of the tasks between the groups. Although not exclusive, the possibility of the latter undercuts inferences that can be made regarding the former; differences in measurement quality (i.e., measurement non-invariance) complicate interpretations of group differences in the constructs. On the other hand, to the extent that there is evidence that the measurement quality of the tasks does *not* differ between groups (i.e., measurement is invariant across groups), differences in observed performance can be taken as stronger evidence of differences in the underlying working memory constructs (Hancock et al., 2009).

Another reason for establishing invariance is to gain a better theoretical understanding of the construct of working memory. That is, even if we never sought to compare the populations, the results of invariance testing will be revelatory, because through invariance testing, we will learn about the structure of bilingual working memory. It will improve our understanding of the working memory model and the generalizability of the construct(s) and the tasks.

Understanding the factor structure of working memory in bilingual children is important in its own right. In addition, understanding the (lack of) invariance in the factor structure between bilingual and monolingual groups is important both for gaining a deeper understanding of the constructs and for facilitating any group comparisons on those constructs or performance on the tasks. To date, there has been no work establishing invariance in working memory models between monolingual and bilingual children. We sought to address this gap in the literature by investigating whether Gray et al.'s (2017) combined working memory model collected from monolingual children was invariant for a Spanish–English bilingual group of second-grade students. We selected this model for two reasons: (1) It is one of the most comprehensive models of working memory in children that has been tested with data and (2) we were able to administer the same tests to our bilingual participants, making the comparison as direct as possible, without the confounds of different measures.

1.2. Working memory models in monolingual English-speaking children

Gray et al. (2017) compared the fit of four different working memory models using 13 tasks with 168 second-grade monolingual English-speaking children: Baddeley's model with the episodic buffer (Baddeley, 2000), Baddeley's model without the episodic buffer (Baddeley & Hitch, 1974), Cowan's model (Cowan, 1988), and a composite of the two models sans episodic buffer. The best fitting model was the composite, three-component model consisting of a Central Executive, Phonological, and Focus of Attention/Visuospatial Factor. This last factor was given this name because all of the visual tasks plus running digit span loaded on this factor and presumably required focused attention because these tasks precluded rehearsal. Here, focus of attention (Cowan, 1988) refers to the use of attention to retain several items in an effortful, limited-capacity manner without the benefit of covert verbal rehearsal; such a maintenance process is thought to occur both for verbal materials under conditions that prevent rehearsal, including running digit span, and for visual materials that are difficult to verbalize (Bunting et al., 2008; Cowan et al., 2005; Gray et al., 2017; Morey & Bieler, 2013).We are not concerned about whether the focus of attention operates by intermittently refreshing each item, thus counteracting a decay process, or by steadily maintaining items in the focus of attention.

In the best-fitting model of Gray et al. (2017), three tasks loaded significantly onto the Central Executive Factor (N-Back Auditory, N-Back Visual[1], and Number Updating), three tasks loaded significantly onto the Phonological Factor (Nonword Repetition, Digit Span, and Phonological binding), and two tasks (Digit Span Running and Cross-Modal binding) loaded weakly on the Phonological Factor. These latter two tasks loaded more strongly (and significantly) onto the Focus of Attention/Visuospatial Factor. The remaining five tasks loaded uniquely and significantly onto the Focus of Attention/Visuospatial Factor (Visual Span Running, Location Span Running, Visual Span, Location Span, and Visual–Spatial Binding).

1.3. Working memory model in Spanish–English bilingual children

The only test of a working memory structural model for bilingual children we are aware of comes from Swanson et al. (2019). Their team used factor analysis to evaluate the fit of one-, two-, and three-factor models of working memory for Spanish–English bilingual children classified as English Language Learners. Model fit was evaluated for cross-sectional groups of 6-, 7-, 8-, 9-, and 10-year-old children for tasks administered in both English and Spanish. While the tasks in each language were designed to be parallel, they were not identical. The three-factor model (Central Executive, Visuospatial Sketchpad, and Phonological Loop), based upon work by Baddeley and Logie (1999), had the best fit across ages and languages, although the fit was not ideal in English for 7-year-olds or in Spanish for 9-year-olds.

In addition to this series of factor analysis models, analyses of measurement invariance were conducted throughout the development for tasks in each language; Swanson et al.'s (2019) research question was focused on the structure of working memory by age. Results provided evidence for invariance up to the level of metric invariance across age groups separately for tasks completed in

[1]In the current manuscript, we refer to these tasks as 'Repetition Detection – Auditory' and 'Repetition Detection – Visual' to more accurately reflect the nature of the tasks. They are the exact same tasks as the Gray et al. (2017) tasks.

https://doi.org/10.1017/S1366728924000580 Published online by Cambridge University Press

English and Spanish. Crucially though, Swanson et al. (2019) did not test for invariance between monolingual and bilingual groups at any age; that was not part of their research question or design. Therefore, their conclusions do not speak to differences in working memory for monolingual English and Spanish–English bilingual children. As noted earlier, establishing measurement invariance between populations is important if one wants to compare mean performance across those populations. In addition, Swanson et al.'s sample was restricted to English learners who do not represent the full range of bilingualism. Hence, although this study makes an important contribution to the literature, questions about Spanish–English invariance of working memory remain.

1.4. Potential differences in working memory performance between monolinguals and bilinguals

Given that no model of working memory has yet been tested for invariance between monolingual and bilingual groups of children, it is difficult to make strong predictions about where differences may occur. However, we can use information about different life experiences in the groups to guide our expectations. In monolingual English-speaking children, the Phonological Factor predicts word learning (Gray et al., 2022), a key component of language. It stands to reason that bilinguals, who manage two languages with phonological differences, may develop or allocate their phonological working memory differently than monolinguals due to their different lived experience. Additionally, based on theories suggesting a bilingual cognitive advantage due to shifting between two languages, central executive function may differ across bilinguals and monolinguals (Arredondo et al., 2017; Bialystok, 2017). However, the literature on central executive benefits in bilingual children is mixed, with many authors finding no bilingual advantage (e.g., Arizmendi et al., 2018; Dick et al., 2019; Duñabeitia et al., 2014). The literature has little to say about what we would expect for the Focus of Attention/Visuospatial Factor in bilingual children, but there is little in terms of lived experience that would suggest that the processing of visual information should differ between groups.

1.5. The current project

We know of only one study that directly tested the structure of working memory in Spanish–English bilingual children (i.e., Swanson et al., 2019). Although this paper makes an important contribution, like any single paper, its contributions were not comprehensive on the topic of bilingual working memory. Understanding bilingual working memory structure is an important first step necessary for theoretical reasons, as well as for developing and evaluating assessments and interventions appropriate for bilinguals. Further, if we are interested in comparing working memory performance between bilingual and monolingual children, evidence for measurement invariance must first be established.

The current project aims to do just that. To establish a model of working memory for bilingual children, our primary research question was:

1) To what extent is the three-factor working memory structure derived by Gray et al. (2017) for monolingual English-speaking second graders a good fit and invariant for bilingual Spanish–English second graders who perform the same tasks?

To the extent that there is between-group invariance, our second research question was:

2) Are there between-group differences in working memory latent factor scores for the monolingual and bilingual groups?

To the extent that measurement invariance is established, and it is appropriate to compare group factor means, we predict that group differences will most likely be observed for the Central Executive and/or Phonological Factors due to the different cognitive and linguistic experiences of bilingual and monolingual children.

2. Methods

2.1. Participants

Recruitment began after IRB approval. A total of 80 Spanish–English bilingual second graders were recruited from schools and community centers in Arizona. We compared their data to 167 monolingual English second graders who were included in Gray et al. (2017). Demographics and performance on assessment for both groups of children can be found in Table 1. We chose second graders because, given that we only had the resources to test enough children to adequately fit the model at one grade, second graders were young enough to see evidence of early working memory development and old enough to accurately ensure children had typical reading skills, which was important to research questions in the larger POWWER study.

The following inclusion criteria applied to all participants: (1) pass a bilateral hearing screening at 1,000, 2,000, and 4,000 Hz; (2) pass a near-visual acuity screening (with correction if needed); (3) pass a color vision screening; (4) enrolled in second

Table 1. Bilingual participant characteristics

Variable	Monolingual		Bilingual	
	M(SD)	Range	M(SD)	Range
Age	7;9 (0;5)	6;10–9;2	7;9 (0;5)	7;0–9;0
MLE	15.39 (1.66)	12–17	12.53 (2.58)	8–17
TOWRE–2	109.45 (8.40)	96–145	108.12 (7.76)	96–127
K-ABC2	117.60 (15.53)	78–160	106.61 (11.77)	80–141
CELF–4	108.75 (9.59)	88–130	93.45 (9.10)	78–117
SCELF–4- Total[a]	–	–	93.48 (11.81)	74–117
SCELF–4 FO	–	–	10.74 (2.37)	6–16
GFTA–2	50.89 (8.54)	7–62	44.80 (10.67)	7–60
EVT–2	112.39 (10.95)	90–137	93.88 (8.88)	77–112
EOWPVT–4: SBE	–	–	109.82 (14.25)	79–145
WRMT	108.23 (9.85)	82–144	102.40 (9.10)	75–121
ADHD[b]	10.19 (8.77)	0–41	7.80 (7.99)	0–38

Note: MLE = mother's level of education; TOWRE-2 = Test of Word Reading Efficiency– Second Edition (Torgesen et al., 2012); K-ABC2 = Kaufman Assessment Battery for Children, Second Edition (Kaufman & Kaufman, 2004); CELF-4 = Clinical Evaluation of Language Fundamentals–Fourth Edition (Semel et al., 2003); SCELF-4 total = Spanish Clinical Evaluations of Language Fundamentals- Fourth Edition standard score (Semel et al., 2006); SCELF-FO = Spanish Clinical Evaluations of Language Fundamentals Fourth Edition–Formulación de Oraciones standard score; GFTA-2 = Goldman-Fristoe Test of Articulation–Second Edition (Goldman & Fristoe, 2000); EVT-2 = Expressive Vocabulary Test–Second Edition (Williams, 2007); EOWPVT-4: SBE = Expressive One-Word Picture Vocabulary Test–4: Spanish-Bilingual Edition standard score (Martin & Brownell, 2012); WRMT = Woodcock Reading Mastery Test, Paragraph Comprehension Subtest-3rd Edition (Woodcock, 2011); ADHD = parental rating of attention-deficit/hyperactivity disorder (ADHD) behaviors using the ADHD Rating Scale–IV Home Version (DuPaul et al., 1998);
[a]Not all children needed to take the entire SCELF.
[b]Lower scores on this measure reflect fewer concerns.

https://doi.org/10.1017/S1366728924000580 Published online by Cambridge University Press

grade at the time of the study; (5) parent reports no history of neuropsychiatric disorders (e.g., ADHD); (6) have a nonverbal IQ standard score of 75 or greater on Kaufman Assessment Battery for Children, Second Edition (Kaufman & Kaufman, 2004); and (6) score above the 30th percentile of on the Goldman–Fristoe Test of Articulation – Second Edition (Goldman & Fristoe, 2000), although scores <30th percentile were accepted if the errors pertained to a single consonant.

The inclusion criteria for bilingual participants were established to ensure children had functional use of both languages and were relatively balanced bilinguals. This was important to ensure that we could potentially see any differences in the Central Executive due to the switching between languages that bilingual children engage in. Children had to be able to hold a conversation in both Spanish and English per the caregiver report, and this skill was verified with direct language testing (reported below). Each child was required to have had at least one Spanish-speaking primary caregiver to ensure adequate exposure to Spanish. The elementary instructional language was required to be English only or English and Spanish.

Simultaneous bilinguals, defined as exposure to English and Spanish before age 3 years, were the majority of the sample. Most of the families (97.44%) provided a home language environment report and the majority reported providing bilingual input by parents (63%) and other relatives (e.g., siblings and grandparents) (30%). Only three families provided monolingual Spanish input (3.9%); those children received English education only (1) or English–Spanish education (2). Given that we focused on children's demonstrated ability in Spanish and English, we did not focus on secondary reports such as percent of language usage, which can be misleading (e.g., parents may not have a sense of a child's use of English at school) to determine eligibility.

Every child had to receive an overall standard score that indicated language development scores within the average range. This could be an 88 or higher on the Clinical Evaluation of Language Fundamentals-Fourth Edition (CELF-4; M = 100; SD = 15; Semel et al., 2003) to demonstrate English skills and a subtest score of 6 or higher on the Formulated Sentence subtest of the CELF-4 Spanish (Semel et al., 2006; M = 10; SD = 3) to demonstrate Spanish skills. Children who received a standard score between 78 and 88 on the CELF-4 (English) were also administered the CELF-4 Spanish in its entirety. As per Barragán et al.'s (2018) analysis of 680 Spanish–English-speaking children from Arizona, a CELF-4 Spanish cut score was set to 78, which has adequate sensitivity (85%) and specificity (80%) for diagnosis of language impairment of Spanish-dominant children. See Table 1 for bilingual participant characteristics and performance on the assessment battery.

2.2. Procedures

Throughout eight 1–1.5-hour sessions, children were administered the test battery described above, as well as working memory tasks from the *Comprehensive Assessment Battery for Children – Working Memory* (CABC-WM; Cabbage et al., 2017). They also completed word learning and executive function tasks, which were not analyzed as part of this study. All research sessions occurred individually with a trained, Spanish–English bilingual research assistant.

2.3. Comprehensive Assessment Battery for Children – Working Memory (Gray et al., n.d.)

Children completed 13 pirate-themed computer-based working tasks while seated 52 cm from a touchscreen computer with a research assistant beside them. All task instructions and stimuli were presented in English and included animations and visuals to facilitate understanding. Task order was randomized across and within testing sessions. Tasks are described below, but please see Gray et al. (2017) Appendix A for even more detailed descriptions of the tasks. Each game began with instructions and practice trials, which the children were required to pass. Children who did not pass training skipped the game and were assigned an imputed score, which was either the lowest average score of children who did pass training or 0 for tasks where the chance was close to 0. To increase motivation, a pirate guide asked for the child's help with each task and provided gold coins and rocks at the end of the task. The coins and rocks did not provide feedback on task performance but instead served as motivation for trying. The child could spend the coins on virtual goods for their pirate after each testing session. Table 2 summarizes children's performance on the battery.

2.3.1. Central executive tasks

These tasks required the storage *and* manipulation of information. The child had to maintain phonological or visual representations in mind while processing incoming stimuli.

Repetition detection – auditory. Robots played individual pure tones (1,000, 1,250, 1,500, 1,750, and 2,000 Hz) to the child who was asked to press a green key on the keyboard if the incoming tone was the same as the previous tone or a red key if the incoming tone was different from the previous tone. The tones were presented by a stationary robot on the screen, followed by a silent period and a screen of only a green square for 3,000 ms, which was the child's cue to respond. Subsequent trials began immediately after the child's response or the elapse of the 3,000 ms, whichever came first. The computer recorded the child's accuracy. Children only judged one tone back for all trials. The dependent variable was overall accuracy. Immediate repetition requires attentional vigilance because items are often repeated in the stream, so that when an item seems familiar, its familiarity could stem from either the just-preceding item or a recent item before that.

Table 2. Descriptive statistics for comprehensive assessment battery for children – working memory

	Monolingual		Bilingual	
Measure	*M*	*SD*	*M*	*SD*
Digit span	19.58	6.84	14.40	5.41
Digit span running	1.85	1.24	1.81	0.89
Location span	10.77	6.11	10.04	5.01
Location span running	1.33	0.66	1.29	0.69
Visual span	6.88	5.76	7.43	6.33
Visual span running	0.88	0.65	0.80	0.69
Auditory repetition detection	0.83	0.16	0.73	0.19
Visual repetition detection	0.74	0.21	0.70	0.22
Nonword repetition	11.54	6.58	8.08	6.20
Number updating	0.87	0.24	0.87	0.22
Cross-modal binding	4.32	2.67	4.18	2.49
Phonological binding span	12.25	6.88	9.23	5.51
Visual binding span	4.40	3.30	4.06	2.75

https://doi.org/10.1017/S1366728924000580 Published online by Cambridge University Press

Repetition detection – visual. Black squares with varying patterns of white dots were shown to the child. The child was asked to press the green key if the incoming square was the same as the previous square or the red key if the incoming square was different. Each square was presented for 1,000 ms followed by a response cue screen for 3,000 ms. Subsequent trials began immediately after the child's response or the elapse of the 3,000 ms, whichever came first. The computer recorded the child's accuracy. Like the repetition detection – auditory task, children only recalled one image back for all trials. The dependent variable was overall accuracy.

Number updating. Children viewed two black-outlined squares on the screen representing two types of toys, each with a single-digit number in them (e.g., 5 2, meaning 5 yoyos and 2 teddy bears) for 2,000 ms. Next, the square outlines would turn red, the digits would disappear, and one of the squares would have an addition operation (e.g., +1, meaning to add 1 to the specified group of toys) for 500 ms. Finally, the square outlines turned green and were void of numbers. The child was expected to remember the numbers, perform the operation of adding 1 to the respective toy box, and say the new numbers, which the RA entered into the computer. The subsequent trial began 50 ms after the RA entered the child's response. The child had to correctly state both numbers to receive a score of 1. If a child responded with an incorrect number but then correctly added the operations to it moving forward, they would receive a score of 0 for the initial incorrect number trial and then a score of 1 for each correctly operated trial.

Digit span running. This task was the same as the digit span task (see below), except for the fact that the digit list varied in length per trial to prevent phonological rehearsal. Digit lists varied from 7 to 10 digits. The child was asked to repeat as many digits as they could remember from the end of the sequence in forward order. We calculated an average of the number of items correctly recalled from the end of the list across all trials.

Running span largely prevents the use of rehearsal processes that can occur with ordinary span tasks, according to several types of evidence. First, unlike span tasks in which the list length is known on most trials and is within the child's capability, in running span there is no primacy effect or enhanced recall of items earlier in the list compared to medial items, a usual signature of rehearsal (Bunting et al., 2008). Second, ordinary and running digit span is equivalent in the prediction of aptitude scores only in children too young to rehearse (Cowan et al., 2005); in older children, standard span scores do not predict aptitudes as well, presumably because verbal rehearsal in older children reduces the strain on attention. Rehearsal is discouraged in running span tasks because the unpredictable endpoint of the list makes it difficult to know which items to rehearse. Most lists are too long to be remembered and recall from the end of the list is required. Consequently, rehearsal of the relevant items would be possible only if the participant with a standard span of N items knew when the last N items were about to occur so that the appropriate rehearsal of these last N items could begin. The alternative possibility of rehearsing starting at the beginning of the list, but then discarding some from the beginning to allow room for more from the end, would be quite challenging to carry out. Therefore, we believe that participants usually accomplish running span tasks by waiting passively for the list to end and then focusing attention on the passive memory of the sensory (or phonological) stimulus stream, providing an index of the efficacy of this attention-based process.

2.3.2. Phonological factor tasks

These tasks measured children's capacity to store and rehearse phonological information.

Digit span. The child was asked to listen to a series of single-syllable, single-digit one-syllable numbers (1–9, excluding 7) presented with one number per second and then repeat as many numbers as they could remember when cued by a green square on the screen. Sequences started with two digits and gradually increased up to eight digits. The child's responses were recorded through their microphone and were entered by the RA. We used weighted scoring, where each completely correct trial was given the value of the length of the trial (e.g., a correct trial of 3 digits was worth 3 points), and all correct trials were added together.

Nonword repetition. After hearing an English-like nonword, the child was asked to repeat it. Their responses were recorded by computer audio. The RA pushed a button to move the task forward after the child's response. The nonword stimuli (16) consisted of four sets per syllable length (two three, four, and five syllables). They contained low-frequency biphones and had no phonological neighbors. Each child's response was transcribed in the lab and given a score of 1 for 100% consonants correct or 0 if not all consonants were correct except for child-specific substitutions seen on their GFTA-2.

2.3.3. Focus of attention/visuospatial tasks

These tasks measured the capacity to store visuospatial information. Unique location tasks and shapes were used to avoid using linguistic resources to name them.

Location span. Trials started with a series of arrows emanating from a black dot in the middle of the screen to one of eight equidistant angle locations around the black dot for 1,000 ms per arrow. The child then saw a screen with the black dot circled by red dots, each representing one of the possible eight locations the arrow pointed to. The child was asked to point to the red dots in the order they saw. The spans started with two arrow sequences and gradually increased up to six. Children earned 1 point for a completely correct trial and 0 points for a trial with any errors. We calculated a weighted score in the same manner as for the digit span task.

Location span running. This task was the same procedure as the location span, but the spans were of random length (five to eight locations) to prevent rehearsal. Children were asked to respond with as many locations from the end of the span in forward order and to press "next" on the screen when finished with their selection. This task was scored in the same way as digit span running, with an average of the number of correct locations identified from the end of the list, across all trials.

Visual span. A series of black polygons were presented individually on the screen for 1,000 ms intervals. After each span, a series of boxes corresponding to the length of the span appeared on the screen with six polygons underneath. The child was asked to touch them in the order they had been presented. Subsequent trials began immediately after the child had filled the boxes. The spans started with one polygon and gradually increased to six. A child received a score of 1 for correctly identifying the shapes and their order and a score of 0 for each span with any errors. This task also used a weighted score.

Visual span running. This task was the same procedure as the visual span, but the span lengths were randomized (i.e., 3 – 6) to prevent rehearsal. The child was asked to respond with as many polygons from the end of the span in forward order and to press "next" on the screen when finished with their selection. This task was scored in the same manner as the previously described running tasks.

2.3.4. Binding tasks

The binding tasks measured working memory capacity when two modalities of information (phonological and/or visuospatial) had to be maintained in working memory to perform well on the task.

https://doi.org/10.1017/S1366728924000580 Published online by Cambridge University Press

Phonological binding span. The child was expected to map English-like nonwords onto non-speech sounds (e.g., beeps and mechanical noises). The child saw a robot that remained on the screen for both sound and nonword. A nonword would be presented 500 ms after participants were presented with a non-speech sound. This continued for spans of one to four words. After a delay of 2,000 ms, the child would hear the sound again and see the speaker icon. The speaker icon and sound would disappear, and a green box would appear, prompting the child to repeat the paired nonword. The research assistant would advance to the next trial.

Nonwords were randomly paired with the non-speech sounds each trial and none were repeated within a trial. The nonwords were single-syllable CVC combinations, each with a low phonotactic probability (7 – 13 phonological neighbors). The child's responses were recorded and later scored by trained phonetic transcribers. A nonword was considered correct if all consonants were correct. A child received a score of 1 for every correct sound – nonword pair and a score of 0 for any errors.

Visuospatial binding span. In each trial, the child was presented with a black polygon in one square on a 4 × 4 grid of 16 squares. The polygon remained for 1,000 ms, followed by a blank grid for 500 ms, then another polygon in a different square for 1,000 ms, repeating this pattern until the span was complete. The spans varied randomly from one to six polygons. The child was prompted to respond with a blank grid and a display of six polygons. They were asked to match the polygons to their locations in the sequence they had appeared. A child received a score of 1 for every correctly identified span, which included the correct polygon – location pair in the correct order. A child received a score of 0 for a span if they had any errors. We calculated a weighted total for this task.

Cross-modal binding. Each trial consisted of a simultaneous presentation of a polygon and nonword, which continued until the span was finished. Spans varied randomly from one to six polygon – nonword pairs. After the span presentation, the child was presented with a nonword and asked to touch the corresponding polygon from a field of 6. The nonwords were played in a different order from their original presentation in the span. Nonwords were single-syllable CVCs that had low phonotactic probabilities, low neighborhood densities, and contained different vowels. These nonwords were distinct from nonwords used in other tasks to prevent interference. A child received a score of 1 for correctly pairing every nonword and polygon in a span and a score of 0 for any errors in the span. We calculated a weighted total for this task.

2.3.5. Factor reliabilities

We evaluated the maximal reliability for the set of working memory tasks with respect to each factor (Bentler, 2007; Hancock & Mueller, 2001; Raykov, 2004) via Coefficient H, which is a function of the standardized loadings for the indicators of a factor (Hancock & Mueller, 2001). Factor reliabilities are a more accurate way to demonstrate reliability compared to individual task reliabilities, as the tasks are not used in isolation, but as indicators of common factors. Hancock's *H* coefficients were computed for each factor separately for each group, using the Excel spreadsheet provided by McNeish (2018). For the Central Executive factor, the estimated H coefficient was 0.68 for the monolingual group and 0.50 for the bilingual group. For the Phonological Loop factor, the estimated H coefficient was 0.55 for the monolingual group and 0.59 for the bilingual group. For the Focus of Attention/Visuospatial factor, the estimated H coefficient was 0.77 for the monolingual group and 0.70 for the bilingual group.

These results provide evidence of comparable reliability across groups except for higher reliability for the monolingual group for the CE factor. At first glance, these values might be considered low. However, we currently lack agreed-upon cutoffs for adequate reliability across assessments (AERA et al., 2014, Chapter 2.) Importantly, these are factor means – not individual means. Procedures to evaluate the reliability of factor means and mean differences have not yet been developed, to our knowledge. Also, our reliabilities are in the range of other analyses of factors in research on working memory in children (e.g., Gathercole et al., 2004; Michalczyk et al., 2013).

2.4. Analytic procedure

We used a Bayesian approach to model estimation for all analyses. This approach was selected for two main reasons. First, prior beliefs about parameter values (e.g., factor loadings) can be incorporated into their estimates, which has the benefit of allowing the restriction of estimates to a range of realistic outcome values (Depaoli, 2021). Second, because the Bayesian framework allows for a more intuitive interpretation of results, communicating beliefs and uncertainty about the parameters directly, avoiding problems of interpretations of frequentist procedures (i.e., confidence intervals and hypothesis tests), which are routinely misinterpreted (Goodman, 2008; Greenland et al., 2016; Jackman, 2009; Morey et al., 2016). For example, a Bayesian 95% highest posterior density interval (HPD) refers to the region of values for which there is a 95% probability that the parameter value is between the limits of that region. This is a direct probability statement about a parameter, conveying the likely values; such an interpretation is often misapplied to the frequentist confidence interval, which does not support such an interpretation (Morey et al., 2016).

RQ 1 – To what extent is the three-factor working memory structure derived by Gray et al. (2017) for monolingual English-speaking second graders a good fit and invariant for bilingual Spanish–English second graders?

We tested the combined factor model of working memory from Gray et al. (2017) at configural, metric, scalar, and strict levels (Putnick & Bornstein, 2016), to assess evidence for measurement invariance between monolingual and bilingual groups. This is done by fitting a series of models with different cross-group constraints and comparing the results of the models.

Our first step was to assess whether there was evidence that observed indicators were associated with the same underlying factor structure (configural invariance) across groups; this model specifies the same pattern of loadings, but not their magnitudes. We then constrained factor loadings to be equal across groups (metric invariance), to assess whether the relationships between the latent factors and observed indicators were the same across groups. Next, we constrained the intercepts to be equal across groups (scalar invariance) to assess whether latent factor means can be compared across groups. Lastly, observed indicator residual error variances were constrained to be equal (strict invariance), to assess whether latent factors accounted for the same amount of variability in observed indicators across groups. For a Bayesian approach to model estimation, it is necessary to specify prior distributions, which quantify the researchers' prior beliefs about parameter values before model estimation and are incorporated with the data to produce a posterior distribution, which expresses beliefs about

https://doi.org/10.1017/S1366728924000580 Published online by Cambridge University Press

the model's parameter values after having observed the data (see Depaoli, 2021 for an accessible, more detailed description). Our prior distribution specifications are summarized in the supplemental material.

2.4.1. Model comparisons

Models were compared using the widely applicable Watanabe-Akaike information criterion (WAIC; Watanabe, 2010), leave-one-out cross-validation (LOO-CV; Vehtari et al., 2017), and Bayes factors (Kass & Raftery, 1995). These criteria evaluate which model (i.e., which level of invariance) is a better predictive tool for future data. For WAIC and LOO-CV, lower values are preferred and differences between models are evaluated based on the ratio of the difference between model expected data and observed data, and its standard error. Bayes factors are interpreted in terms of recommended cut values. Where minimal (i.e., nonmeaningful) differences are present, model comparisons are interpreted in favor of the more parsimonious model.

2.4.2. Model fit

In addition to comparing the models (i.e., levels of invariance) to each other, we also evaluated the fit of each level of invariance individually for whether it is a reasonable summary of the data. We conducted posterior predictive model checking using the marginal (log) likelihood, summarized by the posterior predictive *p*-value (PPP) (Levy, 2011), along with two discrepancy indices: (1) Standardized Root Mean Square Residual (SRMR) (Levy, 2011) and (2) Bayesian Comparative Fit Index (BCFI) (Garnier-Villarreal & Jorgensen, 2020).

This marginal (log) likelihood was selected due to its suitability for sample sizes below 1,000 (Garnier-Villarreal & Jorgensen, 2020). PPP values may range from 0 to 1 with a value of 0.50, suggesting perfect alignment between the model-expected data and observed data (Brooks et al., 1996; Gelman et al., 1996). PPP values much lower than 0.50 are considered an indication that the model does not account for the data well, with 0.10 being the most conservative cutoff for an interpretation that the model is a poor fit to the data (Cain & Zhang, 2019). The SRMR and BCFI used here are conceptually comparable to the frequentist SRMR and CFI, although their calculation produces a distribution rather than a point estimate (Garnier-Villarreal & Jorgensen, 2020; Levy, 2011), which is better suited to a Bayesian interpretation of model fit.

RQ 2 – To the extent that there is between-group invariance, are there between-group differences in latent factor scores for the monolingual and bilingual groups?

2.4.3. Latent factor means

If there is evidence for measurement invariance up to the level of scalar invariance, latent factor means for the bilingual and monolingual groups can be compared for each working memory factor (Depaoli, 2021). Using a Bayesian approach, posterior means, posterior standard deviations, and 95% highest posterior densities (HPD) can be obtained from the posterior distribution of the difference between the factor means for each group. To determine whether between-group differences in latent factor scores are meaningful, the plots of the posterior distributions and the 95% HPDs are evaluated for whether and where zero falls within this interval. If zero falls well within the 95% HPD, then between-group differences are not supported (Kruschke, 2013). To measure the magnitude of the difference Glass's Δ effect sizes were calculated for each latent factor.

2.4.5. Latent factor standard deviations and correlations

In addition to group comparisons for latent factor means, we were also interested in whether there was evidence for between-group differences in factor variability and the relationships between latent factors. To evaluate this, between-group comparisons can be conducted for latent factor standard deviations and correlations. As with latent factor means, a posterior distribution of the difference between latent factor standard deviations and correlations is obtained. Posterior means, posterior standard deviations, and 95% HPDs can be obtained from these posterior distributions.

3. Results

3.1. Model comparisons

For measurement invariance addressed by Research Question 1, the results of pairwise comparisons of models differing in one level of invariance between groups are summarized in Table 3. In each case, model comparison measures were evaluated for whether there was evidence for a preference (i.e., evidence for either model being a better predictive tool for the data) for each level of invariance. The results of the comparisons suggested (a) a preference for the metric invariance model as opposed to the configural invariance model; (b) a negligible difference between the metric and scalar invariance models based on the WAIC and LOO-CV indices, and a strong preference for scalar invariance model based upon the Bayes factor value; and (c) a slight preference for the scalar invariance model over the strict invariance model.

Table 3. Summary of model fit and model comparisons at each level of invariance

	Model comparisons			Model fit		
Level of invariance	WAIC	LOO-CV	BF	Marginal log likelihood (PPP)	SRMR	BCFI
Configural	8,324.14	8,324.65	–	0. 657	0.11 (0.01)	0.99 (0.02)
Metric	8,305.39	8,306.26	−26.99	0. 708	0.11 (0.01)	0.99 (0.02)
Scalar	8,309.44	8,310.55	−21.57	0.500	0.11 (0.01)	0.97 (0.03)
Strict	8,319.62	8,319.82	−1.02	0. 104	0.12 (0.01)	0.90 (0.03)

Note: WAIC = Watanabe-Akaike Information Criterion (Watanabe, 2010); LOO-CV = Leave-One-Out Cross-Validation (Vehtari et al., 2017); BF = Bayes Factor; Negative values favor the model with stronger invariance. BF compares the model in that row to the model in the previous row. PPP = Posterior Predictive *p*-value; SRMR = Standard Root Mean Square Residual (Levy, 2011); BCFI = Bayesian Comparative Fit Index (Garnier-Villarreal & Jorgensen, 2020); SRMR and BCFI are reported in terms of means and standard deviations.

https://doi.org/10.1017/S1366728924000580 Published online by Cambridge University Press

Table 4. Between-group differences in latent factors – posterior distribution summaries

Parameter	Posterior mean	Posterior SD	95% HPD
Glass's Δ effect size			
CE	−0.39	0.21	(−0.81, 0.02)
FoA/V	−0.11	0.18	(−0.47, 0.24)
PL	−1.15	0.25	(−1.68, −0.70)
Latent factor means			
CE	−0.18	0.09	(−0.35, 0.02)
FoA/V	−0.07	0.10	(−0.25, 0.14)
PL	−0.56	0.11	(−0.83, −0.38)
Latent factor standard deviations			
CE	0.08	0.08	(−0.08, 0.24)
FoA/V	0.15	0.09	(−0.02, 0.32)
PL	0.04	0.09	(−0.14, 0.22)
Latent factor correlations			
CE ~ FoA/V	−0.02	0.16	(−0.32, 0.30)
CE ~ PL	−0.17	0.26	(−0.69, 0.35)
FoA/V ~ PL	−0.22	0.21	(−0.63, 0.21)

Note: CE = Central Executive; FoA/V = Focus of Attention/Visuospatial; PL = Phonological; HPD = Highest Posterior Densities; Mean difference and effect size values are interpreted with the monolingual score as the reference value constrained equal to 0. Glass's Δ was calculated by dividing the difference score between the latent variables by the standard deviation from the monolingual group as a reference.

3.1.1. Model Fit

The results of the posterior predictive model checking (summarized in Table 3) indicated that the models performed adequately at the configural, metric, and scalar levels of invariance between groups but not at the strict level. Model fit at each level of invariance based upon the SRMR and BCFI suggested a similar interpretation, with adequate fit at configural, metric, and scalar levels but poorer fit at the strict invariance level. This suggests that the Gray et al. (2017) combined three-factor model of working memory is appropriate for both monolingual and bilingual groups, up to the level of scalar invariance. Collectively, the results suggest the three-factor Gray et al. (2017) model is a sufficient fit for data collected using the CABC–WM (Gray et al., n.d.; summarized by Cabbage et al., 2017) for both monolingual and bilingual groups at the level of scalar invariance. This allows us to make comparisons of differences in latent factor scores for working memory (Hancock et al., 2009).

3.2. *Latent factor means and Glass's Δ effect size*

Relative to the question about the extent of potential between-group differences, posterior distributions for latent factor means were estimated using the monolingual group as a reference group, such that its factor means were set equal to 0, and the bilingual groups' factor means, therefore, capture between-group differences on the latent variables. These between-group differences in factor means are summarized in Table 4, with posterior densities for mean differences and the Glass's Δ effect sizes (Glass et al., 1981) of these differences shown in Figures 1 and 2, respectively.

The largest mean difference was found for the Phonological factor. The posterior for Glass's Δ indicates that, according to the model, the mean for the bilingual group is about 1 standard deviation lower than the mean for the monolingual group and we are 95% certain the difference is between .70 and 1.68 standard deviations. This was the only factor mean difference that did not contain 0 in the HPD. For the Central Executive factor, according to the model, the mean for the bilingual group is estimated to be about .4 standard deviations lower than the mean for the monolingual group; as the 95% interval straddles 0, this expresses that the mean for the bilingual group could be as far as .81 standard deviations below the mean for the monolingual group or it might be trivially different (as the 95% interval extends up to where the bilingual group's mean is .02 larger than the monolingual group's mean). For the Focus of Attention/Visuospatial factor, according to the model, the bilingual group's mean is only about .1 standard deviation below that of the monolingual group, but with considerable uncertainty: the 95% interval suggests this bilingual group's mean ranges from about .47 standard deviations *below* the monolingual group's mean to .24 standard deviations *above* the monolingual group's mean.

3.2.1. Latent factor standard deviations and correlations

In addition to comparing group latent factor means, we also conducted between-group comparisons for latent factor variability (i.e., standard deviations) and relationships between latent factors (i.e., correlations). Posterior means, posterior standard deviations, and 95% HPDs were obtained from the posterior distributions of latent factor standard deviations to compare the latent factor standard deviations between groups. These are summarized in Table 4, with posterior densities shown in Figure 3. The standard deviations for the factors in the bilingual group were slightly smaller than their counterparts in the monolingual group, the largest being for the Focus of Attention/Visuospatial factor, where the difference was estimated to be about .15, with considerable uncertainty (95% HPD from −.02 to .32).

Posterior means, posterior standard deviations, and 95% HPDs were also obtained from the posterior distributions of factor correlations and between-group comparisons. Correlations are summarized in Table 4, with posterior densities shown in Figure 4. For the correlation between Central Executive and Focus of Attention/Visuospatial factors, the values in the groups were trivially different (posterior means of .76 and .74 in the bilingual and monolingual groups, respectively). The differences were slightly larger for the correlation between the Central Executive and Phonological Loop (posterior means of .32 and .16 in the bilingual and monolingual groups, respectively) and for the correlation between the Focus of Attention/Visuospatial and Phonological Loop (posterior means of .50 and .28 in the bilingual and monolingual groups, respectively). For all three correlations, there was considerable uncertainty, such as the 95% HPDs straddle 0, suggesting no strong evidence of group differences in the relationships between the latent factors.

3.2.2. Socioeconomic status as a predictor of latent factors

Some evidence suggests that socioeconomic status may play a role in working memory (e.g., Mooney et al., 2021), specifically that lower SES may be related to lower working memory skills. To investigate whether variance in working memory accounted for by group membership might not otherwise be accounted for by socioeconomic status, we conducted post-hoc analyses adding the mother's level of education as a predictor of each of the latent

https://doi.org/10.1017/S1366728924000580 Published online by Cambridge University Press

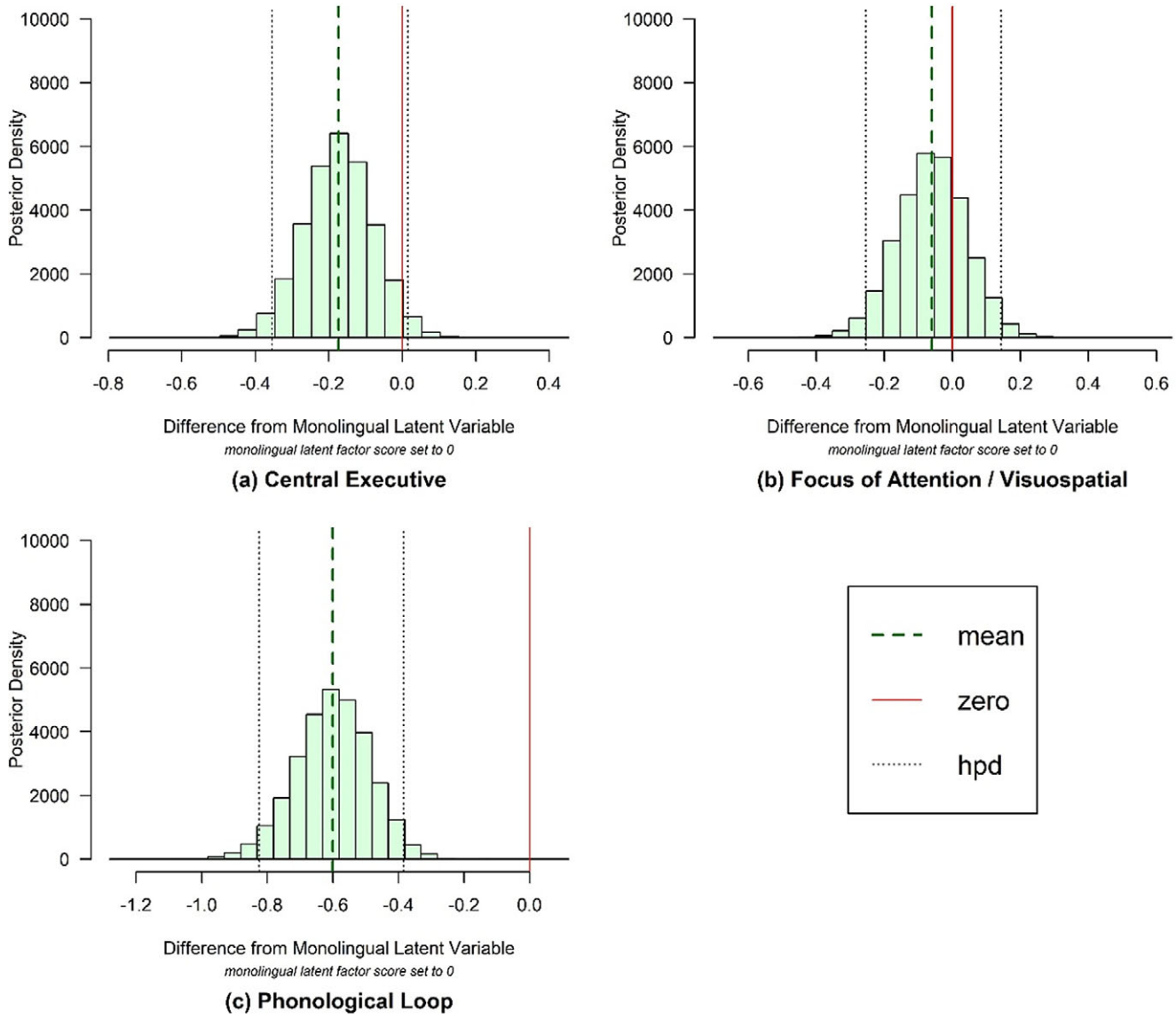

Figure 1. Latent mean difference posterior distribution.
Note: Difference scores are calculated by subtracting monolingual from bilingual latent scores with monolingual mean latent scores set to 0.

factors. Seven cases (two monolingual and five bilingual) were excluded from this analysis as the mother's level of education was not reported.

Posterior means, standard deviations, and 95% HPDs were obtained for latent factor means with the addition regression on years of mother's education (SES) along with Glass's Δ effect sizes (Glass et al., 1981) (summarized in Table 5 and visualized in Figure 5). Results of this analysis suggest that for the Central Executive factor, SES appears to account for a large amount of the between-group differences in latent factor means, with the Glass's Δ effect size dropping from a difference of ~.4 standard deviations between groups to .1 standard deviations, with considerable uncertainty (95% HPD from −2.47 to 2.68). For the Phonological Loop, SES does not appear to account for much difference in latent factor means, with the Glass's Δ effect size of −1.24 (compared to −1.15). However, there is substantial uncertainty around this posterior mean (95% HPD from −4.05 to 1.39), compared to far less uncertainty when SES was not included in the model (95% HPD from −1.68 to .70). Notably, the inclusion of SES in the model appears to suggest that the groups do differ in their Focus of Attention/Visuospatial factor means, with the bilingual group having a mean that is 1.64 standard deviations above that of the monolingual group. However, as with the other two factors, there is considerable uncertainty around this posterior mean (95% HPD from −.85 to 4.05). Although this suggests some evidence that accounting for SES alters how groups differ in these factors, the presence of considerable uncertainty limits the strength of the conclusions that can be drawn.

4. Discussion

There is a pressing need to understand the theoretical structure of working memory in Spanish–English bilingual children, yet currently there is only one published model (Swanson et al., 2019). To understand whether structure varies for monolingual English and

https://doi.org/10.1017/S1366728924000580 Published online by Cambridge University Press

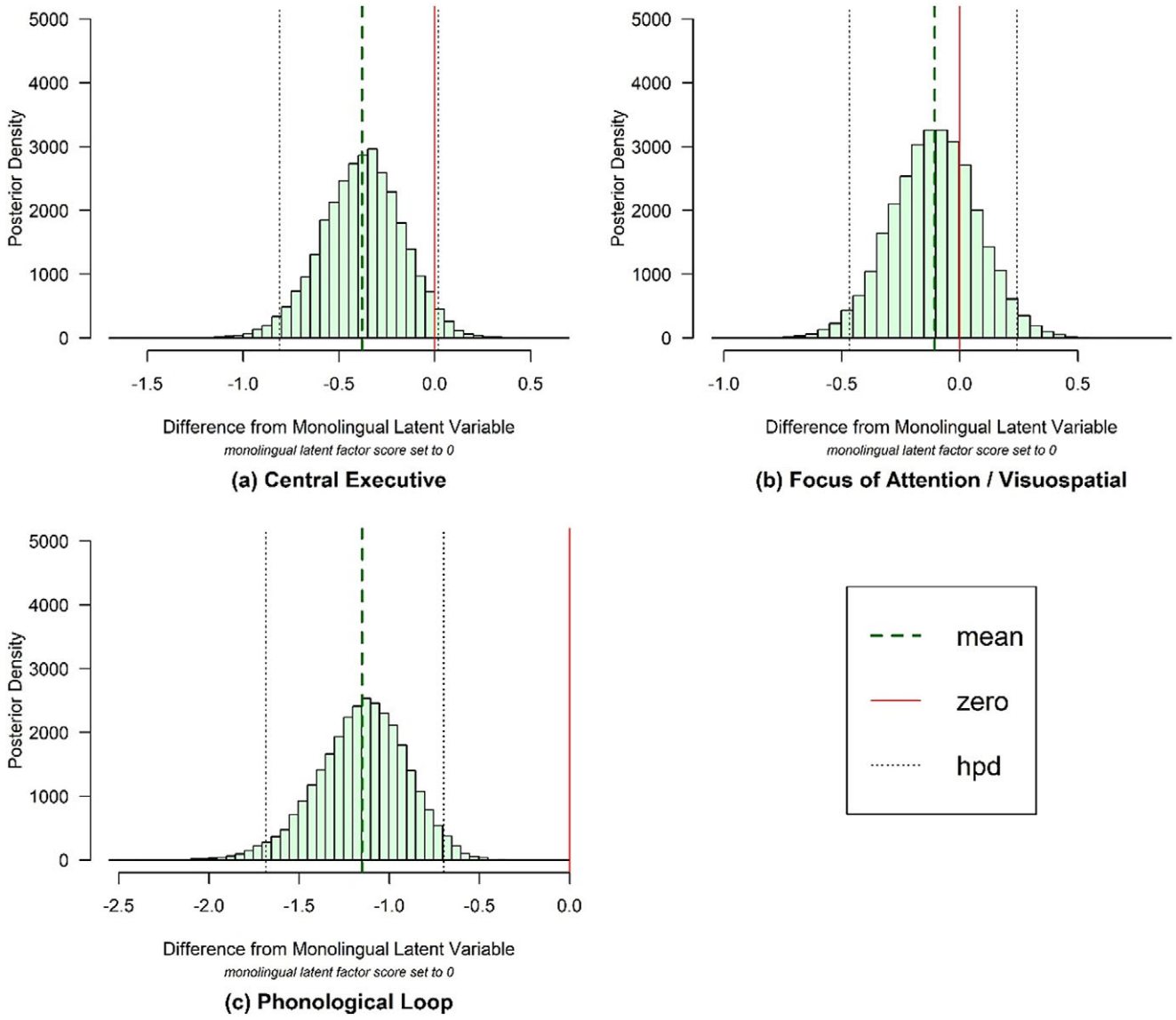

Figure 2. Posterior densities for latent factor score difference Glass's Δ effect sizes.

bilingual Spanish–English-speaking children, we measured the invariance of a preferred model across groups. Our results revealed appropriate fit and invariance between the Gray et al. (2017) monolingual English three-factor combined model (i.e., Central Executive, Phonological, and Focus of Attention/Visuospatial factors) for bilingual Spanish–English and monolingual English typically developing second graders. Despite differences in the cognitive and linguistic demands of bilingual versus monolingual experiences, the structure of working memory appears to be similar across groups.

Our findings generally compare well to the one existing model of bilingual working memory from Swanson et al. (2019), in that we both found that a three-factor model with similar components (e.g., central executive, phonology, and visuospatial) was a good fit. This is despite the differences in our populations and tasks. For example, we studied a single grade, while Swanson et al.'s (2019) data suggested that there may be differences in model fit across ages. Also, Swanson et al. tested English Learners, while most of our children were simultaneous bilinguals with strong English language skills. Like Swanson et al., our participants were primarily from the Southwest. Different types of bilingual experiences could lead to different outcomes. Another difference between our studies is that we used the same tasks with our monolingual and bilingual participants to conduct measurement invariance testing and were able to establish evidence that the same model specification (i.e., Gray et al., 2017) is a good fit for our bilingual participants. This is important because we cannot take for granted that different tasks have comparable factor structures. While there are still many questions, these two quite different approaches to learning about bilingual working memory converge upon the notion of a three-factor model.

One implication of this finding has to do with future educational design. Working memory has not yet been fully leveraged in intervention design, despite its potential to address current stagnation in academic performance levels across the United States (e.g., NAEP scores in math and reading). For example, knowledge of a student's working memory profile allows for the design of instructional or intervention approaches that take advantage of or compensate for those profiles (e.g., engaging in effortful retrieval or

https://doi.org/10.1017/S1366728924000580 Published online by Cambridge University Press

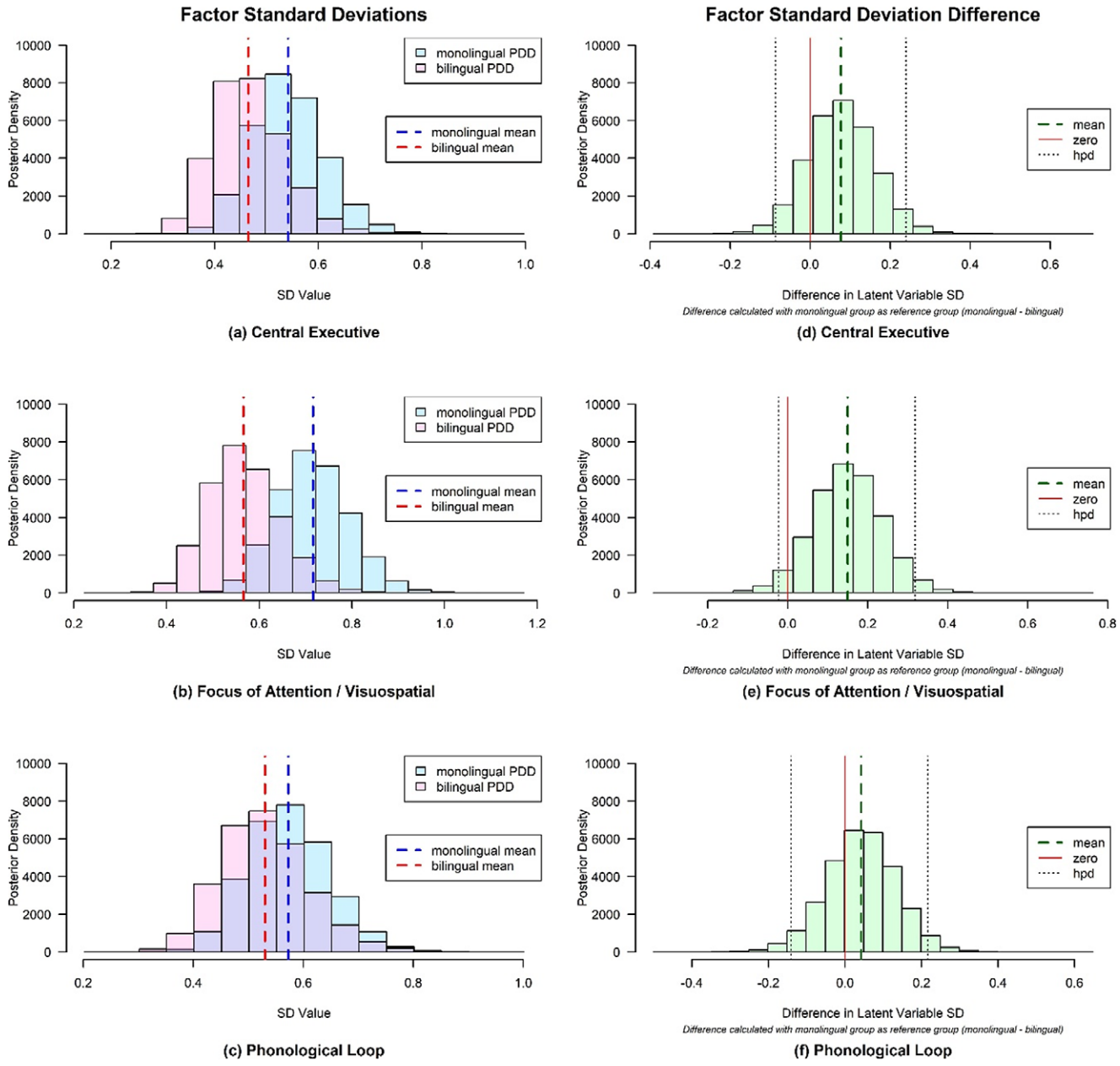

Figure 3. Factor standard deviation posterior distributions.

errorless learning supports). Understanding bilingual working memory means that when educators begin to design interventions to support working memory skills in students (e.g., Colmar et al., 2020; Swanson et al., 2015), they will have a sense of whether the interventions might be appropriate for monolingual students, bilingual students, or both. None of this work can begin in an informed way until we understand the structure of bilingual working memory.

In terms of differences in levels of performance between groups, we focus on the results that incorporate SES into the analyses given the evidence that SES influences working memory (Mooney et al., 2021) and because of the differences in SES between our groups. It is difficult to interpret any between-group differences with confidence due to the high levels of uncertainty associated with the findings; however, we can view the results through a theoretical lens.

Recall that there was nearly no difference between the groups relative to the Central Executive (roughly .1 SD difference), although the HPD ranged from −2.5 to +2.68, meaning that we are 95% certain that it might be equally likely that either group has a meaningful advantage or disadvantage. This lack of clarity squares well with the mixed findings within the literature. There is no reason to assume that one group of children with typical development would have better cognitive skills than another, although evidence points to the fact that experience can influence brain structure and function (e.g., Arredondo et al., 2017; Bialystok, 2017), thus bilingual and monolingual children might show some differences. While some might predict a Central Executive advantage for bilingual children, more recent evidence casts doubt on whether there is a real bilingual advantage (e.g., Arizmendi et al., 2018; Dick et al., 2019; Duñabeitia et al., 2014), with Lowe et al.'s (2021) meta-analysis of executive

https://doi.org/10.1017/S1366728924000580 Published online by Cambridge University Press

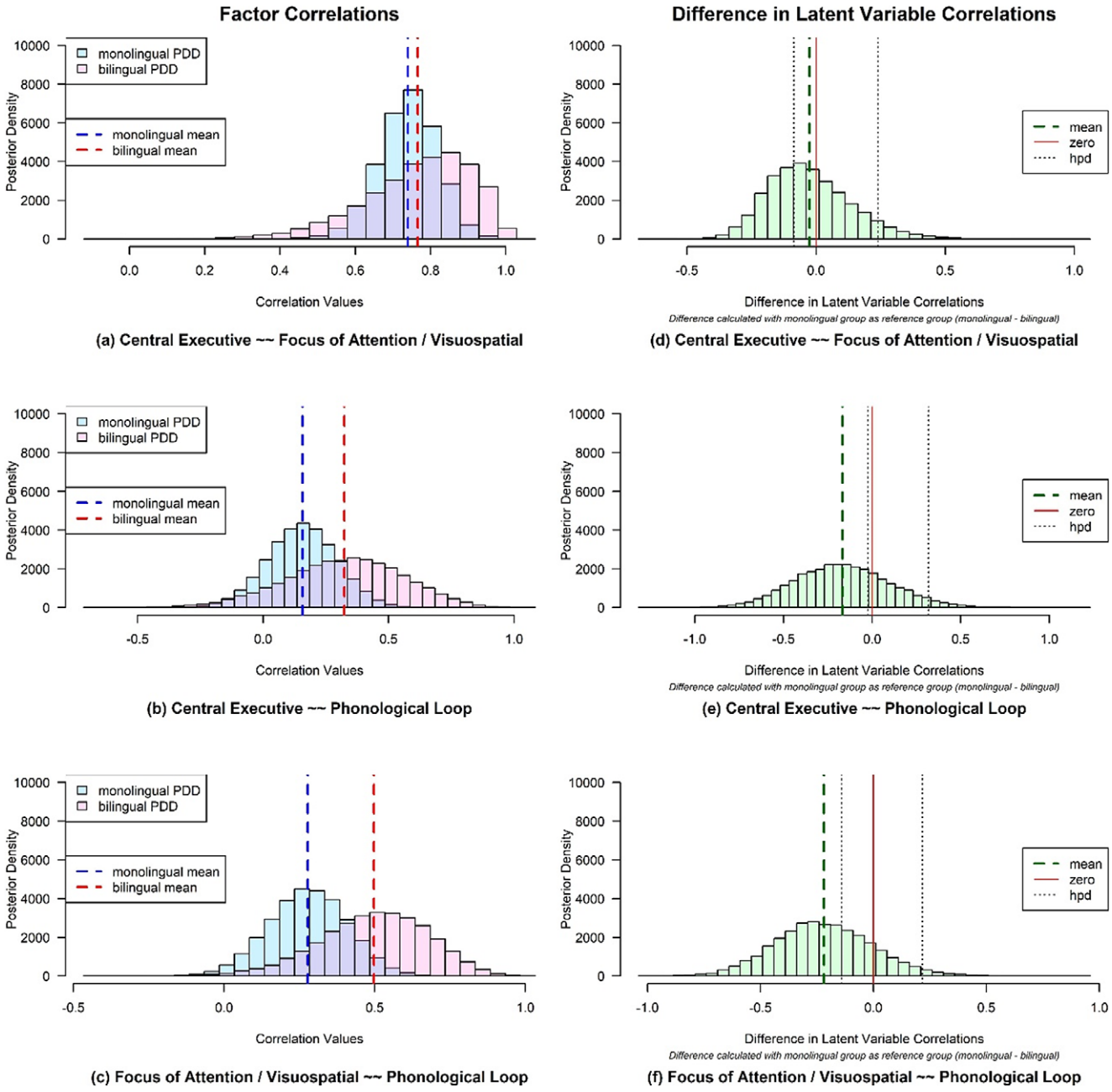

Figure 4. Factor correlation posterior distributions.

Table 5. Between-group differences in latent factors with years of mother education – posterior distribution summaries

Parameter	Posterior mean	Posterior SD	95% HPD
Latent factor means			
CE	0.05	0.61	(−1.20, 1.22)
FoA/V	0.95	0.71	(−0.48, 2.32)
PL	−0.65	0.72	(−2.07, 0.77)
Glass' Δ effect size			
CE	0.12	1.31	(−2.47, 2.68)
FoA/V	1.64	1.24	(−0.81, 4.05)
PL	−1.24	1.38	(−4.05, 1.39)

Note: CE = Central Executive; FoA/V = Focus of Attention/Visuospatial; PL = Phonological; HPD = Highest Posterior Densities; Mean difference and effect size values are interpreted with the monolingual score as the reference value constrained equal to 0.

function in bilingual children describing any potential advantage as small and variable.

Even with SES factored in, we did find a difference favoring the monolingual children in the Phonological Loop. However, there is still considerable uncertainty (i.e., HPD of roughly −4 to +1.4) associated with this finding. Theoretically, a difference favoring the monolingual group makes sense given the monolingual testing. Bilingual children tend to have less experience with the phonology of any one language compared to a monolingual child. For example, the children in this study were also part of a word-learning study. When learning new words, the bilingual children performed equivalently to the monolingual children on sound sequences that were shared across Spanish and English but were less accurate on sound sequences that only occurred in English (Erikson et al., 2021). By using English phonology in our working memory tasks, we may not have assessed the full set of skills for the bilingual group

https://doi.org/10.1017/S1366728924000580 Published online by Cambridge University Press

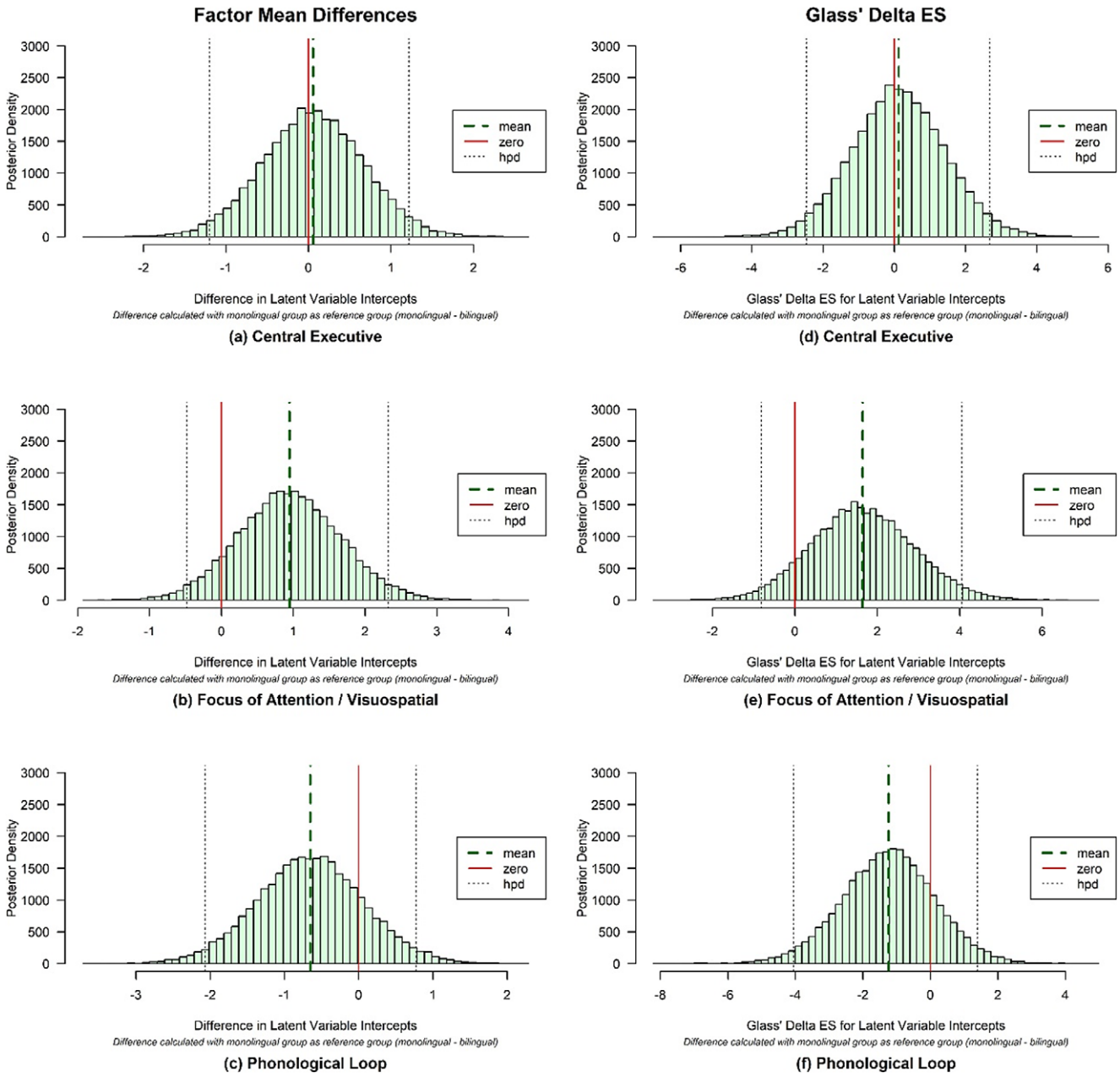

Figure 5. Differences in latent factors accounting for years of mother education – posterior distributions.

and thus the level differences may not hold if a fully bilingual assessment were administered. To date, we are unaware of any working memory assessments that are designed to be administered bilingually, but we are working on this kind of measure.

The between-group finding that was least expected was the bilingual advantage for the Focus of Attention/Visuospatial component. Again, there was considerable uncertainty (i.e., HPD of roughly −0.85 to +4) for this finding. The literature would not lead us to assume that being bilingual would lead to a visual processing advantage. One possibility that might explain this finding is that many visual tasks are associated (perhaps unintentionally) with phonological components. For example, when a child is asked to remember familiar shapes on a memory task, it is difficult to inhibit using the lexical labels for those items (e.g., square and circle). By intentionally creating tasks in our battery that limited the use of lexical labels (e.g., creating difficult-to-name polygons), we may have lessened the phonological burden and allowed a previously undetected skill in bilingual children to emerge. Overall, though, these between-group findings should be interpreted within the context of the high level of uncertainty surrounding each finding.

4.1. Limitations

We recognize that our inferences are based on the interpretation of statistical models that attempt to simplify the complexity of real-world phenomena (Mislevy, 2018), and that the use of such models to facilitate inference is subject to myriad types of errors (Little et al., 2017). For example, the inference of scalar invariance between bilingual and monolingual groups for the working memory factor structure is conditional on decisions made in formulating and using

https://doi.org/10.1017/S1366728924000580 Published online by Cambridge University Press

the model, in terms of construct definition, task design, data collection, and data analysis. Similarly, the inferences regarding group differences in average levels of the Phonological and Focus of Attention/Visuospatial factor (and no group differences in the average levels for the Central Executive factor) are model-based and conditional on all these inputs and steps that led to the use of these models. Thus, our inferences regarding group differences (or the lack thereof) should be seen as being made through a model, with attending caution considering any model to be a necessarily simplified lens on a more complex and nuanced real-world situation. Future work may, and should, challenge these inferences based on things like analyses of other data sets, alternative lenses on the constructs and measurement of working memory, or even the framing of bilingual and monolingual populations (e.g., examining variation in bilingual experiences).

5. Summary

Our study provided evidence that the structure of working memory does not differ between monolingual English-speaking second graders in the southwestern United States and their Spanish–English bilingual peers. There is evidence to suggest that the structure of working memory may change with age and perhaps with the number and types of heritage languages spoken. It is important to test these hypotheses in future studies to provide both a theoretical and practical foundation for researchers and practitioners seeking to study working memory development and to design assessments, curricula, and interventions to help children who have working memory deficits.

Data availability. The data and codebook for this study can be accessed through LDBase. https://www.ldbase.org/datasets/a811b0d9-5f67-4c2e-ab14-6040e0acf9c1 (DOI: 10.33009/ldbase.1680030967.159b). Access to the study materials is available on request through: https://redcap.rc.asu.edu/surveys/?s=PM8X3ATA9Y. Analysis files are available upon request from the authors.

Acknowledgments. *In memory of our late colleague Samuel (Sam) Green. We gratefully acknowledge his valuable contributions to this research. This work was supported by the National Institute on Deafness and Other Communication Disorders Grant R01 DC010784, awarded to Shelley Gray. We are deeply grateful to our research team and all the school administrators, teachers, children, and families who participated. Key personnel included (in alphabetical order) Shara Brinkley, Gary Carstensen, Cecilia Figueroa, Karen Guilmette, Trudy Kuo, Bjorg LeSueur, Annelise Pesch, and Jean Zimmer. Many students also contributed to this work including (in alphabetical order) Lauren Baron, Alexander Brown, Jessie Erikson, Nora Schlesinger, Nisha Talanki, and Hui-Chun Yang.

Competing interest. Mary Alt and Genesis Arizmendi are employed as faculty at University of Arizona. Kimberly Leon and Sarah Lynn Neiling are doctoral students at University of Arizona and receive student financial support. Roy Levy and Shelley Gray are employed as faculty at Arizona State University and DeAnne Hunter is a doctoral student there, who received student financial support. Nelson Cowan is employed as faculty at University of Missouri. The authors have received grant support from the NIH, OSERP/OSEP, and the NSF.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association

Ahmed, S. F., Tang, S., Waters, N. E., & **Davis-Kean, P.** (2019). Executive function and academic achievement: Longitudinal relations from early childhood to adolescence. *Journal of Educational Psychology*, **111**(3), 446–458. https://doi.org/10.1037/edu0000296

Arizmendi, G. D., Alt, M., Gray, S., Hogan, T. P., Green, S., & **Cowan, N.** (2018). Do bilingual children have an executive function advantage? Results from inhibition, shifting, and updating tasks. *Language, Speech, and Hearing Services in Schools*, **49**(3), 356–378. https://doi.org/10.1044/2018_LSHSS-17-0107

Arredondo, M. M., Hu, X. S., Satterfield, T., & **Kovelman, I.** (2017). Bilingualism alters children's frontal lobe functioning for attentional control. *Developmental Science*, **20**(3), e12377. https://doi.org/10.1111/desc.12377

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, **4**, 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. D., & **Hitch, G.** (1974). Working memory. In G. A. Bower (Ed.), *Recent advances in learning and motivation* (Vol. **8**, pp. 47–90). Academic Press.

Baddeley, A. D., & **Logie, R. H.** (1999). The multiple-component model. In A. Miyake, & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28–61). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.005

Barragán, B., Castilla-Earls, A., Martinez-Nieto, L., Restrepo, M. A., & **Gray, S.** (2018). Performance of low-income dual language learners attending English-only schools on the clinical evaluation of language fundamentals-fourth edition, Spanish. *Language, Speech, and Hearing Services in Schools*, **49**(2), 292–305. https://doi.org/10.1044/2017_LSHSS-17-0013

Bentler, P. M. (2007). Covariance structure models for maximal reliability of unit-weighted composites. In S. Lee (Ed.), *Handbook of computing and statistics with applications* (Vol. **1**, pp. 1–19). Elsevier. https://doi.org/10.1016/S1871-0301(06)01001-8

Bialystok, E. (2017). The bilingual adaptation: How minds accommodate experience. *Psychological Bulletin*, **143**(3), 233–262. https://doi.org/10.1037/bul0000099

Brooks, S., Gelman, A., Carlin, J. B., Stern, H. S., & **Rubin, D. B.** (1996). Bayesian data analysis. In *The statistician* (Vol. **45**, Issue 2). https://doi.org/10.2307/2988417

Bunting, M. F., Cowan, N., & **Colflesh, G. H.** (2008). The deployment of attention in short-term memory tasks: Tradeoffs between immediate and delayed deployment. *Memory & Cognition*, **36**, 799–812. https://doi.org/10.3758/mc.36.4.799

Cabbage, K. L., Brinkley, S., Gray, S., Alt., M., Cowan, N., Green, S., Kuo, T., & **Hogan, T.** (2017). Assessing working memory in children: The comprehensive assessment battery for children – working memory (CABC-WM). *Journal of Visualized Experiments (JoVE)*, **124**, e55121. https://doi.org/10.3791/55121

Cain, M. K., & **Zhang, Z.** (2019). Fit for a Bayesian: An evaluation of PPP and DIC for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, **26**(1), 39–50. https://doi.org/10.1080/10705511.2018.1490648

Colmar, S., Double, K., Davis, N., Sheldon, L., Phillips, N., Cheng, M., & **Briddon, S.** (2020). Memory mates: An evaluation of a classroom-based, student-focused working memory intervention. *Journal of Psychologists and Counsellors in Schools*, **30**(2), 159–171. https://doi.org/10.1017/jgc.2020.9

Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. *Educational Psychology Review*, **26**, 197–223. https://doi.org/10.1007/s10648-013-9246-y

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin*, **104**, 163–191. https://doi.org/10.1037/0033-2909.104.2.163

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & **Conway, A.R.A.** (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, **51**, 42–100. doi: 10.1016/j.cogpsych.2004.12.001

Cowan, N. (2022). Working memory development: A 50-year assessment of research and underlying theories. *Cognition*, **224**, 105075. https://doi.org/10.1016/j.cognition.2022.105075

Depaoli S. (2021). *Bayesian structural equation modeling*. Guilford Press.

Dick, A. S., Garcia, N. L., Pruden, S. M., Thompson, W. K., Hawes, S. W., Sutherland, M. T., … & **Gonzalez, R.** (2019). No evidence for a bilingual executive function advantage in the ABCD study. *Nature Human Behaviour*, **3**(7), 692–701. https://doi.org/10.1038/s41562-019-0609-3

Duñabeitia, J. A., **Hernández, J. A.**, **Antón, E.**, **Macizo, P.**, **Estévez, A.**, **Fuentes, L. J.**, & **Carreiras, M.** (2014). The inhibitory advantage in bilingual children revisited: Myth or reality? *Experimental Psychology*, **61**(3), 234. https://doi.org/10.1027/1618-3169/a000243

DuPaul, G. J., **Power, T. J.**, **Anastopoulos, A. D.**, & **Reid, R.** (1998). *ADHD rating scale IV: Checklists, norms, and clinical interpretations*. Guilford Press.

Erikson, J. A., **Alt, M.**, **Gray, S.**, **Green, S.**, **Hogan, T. P.**, & **Cowan, N.** (2021). Phonological vulnerability for school-aged Spanish-English-speaking bilingual children. *International Journal of Bilingual Education and Bilingualism*, **24**(5), 736–756. https://doi.org/10.1080/13670050.2018.1510892

Garnier-Villarreal, M., & **Jorgensen, T. D.** (2020). Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods*, **25**(1), 46–70. https://doi.org/10.1037/met0000224

Gathercole, S. E., **Pickering, S. J.**, **Ambridge, B.**, & **Wearing, H.** (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, **40**(2), 177. https://doi.org/10.1037/0012-1649.40.2.177

Gelman, A., **Meng, X.-L.**, & **Stern, H.** (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Academia Sinica*, **6**(4), 733–760. https://www.jstor.org/stable/24306036

Glass, G. V., **McGaw, B.**, & **Smith, M. L.** (1981). *Meta-analysis in social research* (Vol. **56**). Beverly Hills, CA: Sage.

Goldman, R., & **Fristoe, M.** (2000). *Goldman-fristoe test of articulation – 2*. Pearson.

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, **45**(3), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Gray, S., **Green, S.**, **Alt, M.**, **Hogan, T.**, **Kuo, T.**, **Brinkley, S.**, & **Cowan, N.** (2017). The structure of working memory in young children and its relation to intelligence. *Journal of Memory and Language*, **92**, 183–201. https://doi.org/10.1016/j.jml.2016.06.004

Gray, S. I., **Levy, R.**, **Alt, M.**, **Hogan, T. P.**, & **Cowan, N.** (2022). Working memory predicts new word learning over and above existing vocabulary and nonverbal IQ. *Journal of Speech, Language, and Hearing Research*, **63**, 216–233. https://doi.org/10.1044/2021_JSLHR-21-00397

Greenland, S., **Senn, S. J.**, **Rothman, K. J.**, **Carlin, J. B.**, **Poole, C.**, **Goodman, S. N.**, & **Altman, D. G.** (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, **31**(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Grundy, J. G., & **Timmer, K.** (2017). Bilingualism and working memory capacity: A comprehensive meta-analysis. *Second Language Research*, **33** (3), 325–340. https://doi.org/10.1177/0267658316678286

Gunnerud, H. L., **Ten Braak, D.**, **Reikerås, E. K. L.**, **Donolato, E.**, & **Melby-Lervåg, M.** (2020). Is bilingualism related to a cognitive advantage in children? A systematic review and meta-analysis. *Psychological Bulletin*, **146**(12), 1059–1083. https://doi.org/10.1037/bul0000301

Hancock, G. R., & **Mueller, R. O.** (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future – A festschrift in honor of Karl Jöreskog* (pp. 195–216). Scientific Software International.

Hancock, G. R., **Stapleton, L. M.**, & **Arnold-Berkovits, I.** (2009). The tenuousness of invariance tests within multisample covariance and mean structure models. In T. Teo, & M. S. Khine (Eds.), *Structural equation modeling: Concepts and applications in educational research (pp.* 137–174*).* Sense Publishers.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. Wiley.

Kass, R. E., & **Raftery, A. E.** (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Kaufman, A. S., & **Kaufman, N. L.** (2004). *Kaufman assessment battery for children* (2nd ed.).Pearson.

Kruschke, J.K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, **142**(2)573–603. https://doi.org/10.1037/a0029146

Levy, R. (2011). Bayesian data-model fit assessment for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, **18**(4), 663–685. https://doi.org/10.1080/10705511.2011.607723

Little, T. D., **Widaman, K. F.**, **Levy, R.**, **Rodgers, J. L.**, & **Hancock, G. R.** (2017). Error, error in my model, who's the fairest error of them all? *Research in Human Development*, **14**(4), 271–286. https://doi.org/10.1080/15427609.2017.1370965

Lowe, C.J., **Cho, I.**, **Goldsmith, S.F.**, & **Morton, J.B.** (2021). The bilingual advantage in children's executive functioning is not related to language status: A meta-analytic review. *Psychological Science*, **32**, 1115–1146. https://doi.org/10.1177/0956797621993108

Martin, N., & **Brownell, R.** (2012). *Expressive one-word picture vocabulary Test – 4: Spanish-bilingual edition*. Academic Therapy Publications.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, **23**(3), 412–433. https://doi.org/10.1037/met0000144

Michalczyk, K., **Malstädt, N.**, **Worgt, M.**, **Könen, T.**, & **Hasselhorn, M.** (2013). Age differences and measurement invariance of working memory in 5-to 12-year-old children. *European Journal of Psychological Assessment*, **29**, 220–229. https://doi.org/10.1027/1015-5759/a000149

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Monnier, C., **Boiché, J.**, **Armandon, P.**, **Baudoin, S.**, & **Bellocchi, S.** (2022). Is bilingualism associated with better working memory capacity? A meta-analysis. *International Journal of Bilingual Education and Bilingualism*, **25** (6), 2229–2255. https://doi.org/10.1080/13670050.2021.1908220

Mooney, K. E., **Prady, S. L.**, **Barker, M. M.**, **Pickett, K. E.**, & **Waterman, A. H.** (2021). The association between socioeconomic disadvantage and children's working memory abilities: A systematic review and meta-analysis. *Plos One*, **16**(12), e0260788. https://doi.org/10.1371/journal.pone.0260788

Morey, C.C., & **Bieler, M.** (2013). Visual short-term memory always requires attention. *Psychonomic Bulletin & Review*, **20**, 163–170. https://doi.org/10.3758/s13423-012-0313-z

Morey, R. D., **Hoekstra, R.**, **Rouder, J. N.**, **Lee, M. D.**, & **Wagenmakers, E. J.** (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, **23**(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Peng, P., & **Kievit, R. A.** (2020). The development of academic achievement and cognitive abilities: A bidirectional perspective. *Child Development Perspectives*, **14**(1), 15–20. https://doi.org/10.1111/cdep.12352

Putnick, D. L., & **Bornstein, M. H.** (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, **41**, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy*, **35**, 299–331. https://doi.org/10.1016/S0005-7894(04)80041-8

Semel, E., **Wiig, E. H.**, & **Secord, W. A.** (2003). *Clinical evaluation of language fundamentals* (4th ed.). Pearson.

Semel, E., **Wiig, E. H.**, & **Secord, W. A.** (2006). *Clinical evaluation of language fundamentals–4th edition, Spanish*. The Psychological Corporation.

Simone, A.N., **Marks, D.K.**, **Bédard, A. C.**, & **Halperin, J. M.** (2018). Low working memory rather than ADHD symptoms predicts poor academic achievement in school-aged children. *Journal of Abnormal Child Psychology*, **46**(2), 277–290. https://doi.org/10.1007/s10802-017-0288-3

Swanson, H. L., **Arizmendi, G. D.**, & **Li, J. T.** (2021). Working memory growth predicts mathematical problem-solving growth among emergent bilingual children. *Journal of Experimental Child Psychology*, **201**, 104988. https://doi.org/10.1016/j.jecp.2020.104988

Swanson, H. L, **Kudo, M. F.**, & **Van Horn, L.** (2019). Does the structure of working memory in EL children vary across age and two language systems? *Memory*, **27**(2), 174–191. https://doi.org/10.1080/09658211.2018.1496264

Swanson, H. L., **Lussier, C. M.**, & **Orosco, M. J.** (2015). Cognitive strategies, working memory, and growth in word problem solving in children with math difficulties. *Journal of Learning Disabilities*, **48**(4), 339–358. https://doi.org/10.1177/0022219413498771

Torgesen, J., **Wagner, R.**, & **Rashotte, C.** (2012). *Test of word reading efficiency* (2nd ed.). Pro-Ed.

Vehtari, A., **Gelman, A.**, & **Gabry, J.** (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, **27**(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely application information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**, 3571–3594.

Williams, K. T. (2007). *Expressive vocabulary test* (2nd ed.). Pearson.

Woodcock, R. W. (2011). *Woodcock reading mastery test*. Pearson.

https://doi.org/10.1017/S1366728924000580 Published online by Cambridge University Press