Journal of the International Neuropsychological Society (2021), **27**, 825–834

doi:10.1017/S1355617720001265

Exploring the Factor Structure of the NIH Toolbox Cognition Battery in a Large Sample of 8-Year-Old Children in Aotearoa New Zealand

Denise Neumann[1,2,*], Elizabeth R. Peterson[1,2], Lisa Underwood[2,3], Susan M.B. Morton[2,3] and Karen E. Waldie[1,2]

[1]School of Psychology, the University of Auckland, Auckland, New Zealand
[2]Centre for Longitudinal Research – He Ara ki Mua, the University of Auckland, Auckland, New Zealand
[3]School of Population Health, the University of Auckland, Auckland, New Zealand

(Received June 23, 2020; Final revision October 29, 2020; Accepted October 30, 2020; First Published Online January 11, 2021)

Abstract

Objective: The objective of this study was to derive a factor structure of the measures of the National Institutes of Health (NIH) Toolbox Cognition Battery (CB) that is representative of cognitive abilities in a large ethnically diverse cohort of 8-year-old children in Aotearoa New Zealand. Methods: Our sample comprised of 4298 8-year-old children from the *Growing Up in New Zealand* study. We conducted exploratory and confirmatory factor analysis for the NIH Toolbox CB measures to discover the best-fitting factor structure in our sample. Measurement invariance of the identified model was tested across child's gender, socio-economic status (SES), and ethnicity. Results: A three-dimensional factor structure was identified, with one factor of Crystallised Cognition (Reading and Vocabulary), and two distinguished factors of fluid cognition: Fluid Cognition I (Attention/Inhibitory Control, Processing Speed, and Cognitive Flexibility) and Fluid Cognition II (Working Memory, Episodic Memory). The results demonstrate excellent model fit, but reliability of the factors was low. Measurement invariance was confirmed for child's gender. We found configural, but neither metric nor scalar, invariance across SES and the four major ethnic groups: European, Māori, Pacific Peoples, and Asian. Conclusion: Our findings show that, at the age of 8 years, fluid abilities are more strongly associated with one another than with crystallised abilities and that fluid abilities need to be further differentiated. This dimensional structure allows for comparisons across child's gender, but evaluations across SES and ethnicity within the Aotearoa New Zealand context must be conducted with caution. We recommend using raw scores of the individual NIH Toolbox CB measures in future research.

Keywords: Growing up in New Zealand, Longitudinal, Cognition, Language, Fluid, Crystallised

INTRODUCTION

The cognition domain, measured by the National Institutes of Health (NIH) Toolbox Cognition Battery (CB) of the NIH Toolbox® for Assessment of Neurological and Behavioral Function was developed as a standard set of measures of cognitive function across the lifespan (aged 3–85 years), intended to address the needs for a brief assessment tool for large-scale epidemiologic and longitudinal studies and to allow for international cross-study comparisons (Gershon et al., 2010). Six subdomains are assessed by the NIH Toolbox CB: executive function (with tests of cognitive flexibility and inhibitory control/attention), episodic memory, language, reading, working memory, and processing speed (Weintraub, Bauer, et al., 2013).

In addition to the scores of the individual measures for these subdomains, the NIH Toolbox CB provides composite scores, aiming to allow for evaluation of overall and higher level cognitive functioning. The basis for defining the NIH Toolbox CB composite scores is to group those subtests together as a composite score that relate to each other and share theoretical and psychometric characteristics across the lifespan (Akshoomoff et al., 2013). Basis for this approach is one of the most popular models for the structure of human intelligence, the two-component theory of intellectual development (Cattell, 1971; Horn, 1970), which postulates the organisation of cognitive abilities as indicators of *fluid* and *crystallised* intelligence (Li et al., 2004). Fluid cognitive abilities are defined as problem-solving and

*Correspondence and reprint requests to: Denise Neumann, PhD, School of Psychology, Faculty of Science, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand. E-mail: d.neumann@auckland.ac.nz

information processing and play an important role in adapting to novel situations in everyday life. Fluid abilities are less dependent on past learning experiences and cultural biases but depend more on biological processes. Crystallised cognitive abilities represent verbal knowledge and skills, depend upon past learning experiences, and are strongly influenced by education and cultural exposure (Heaton et al., 2014; Li et al., 2004). As first discovered by Spearman (1961), all cognitive tests are positively correlated with each other, encompassing one common factor, the general factor of intelligence (Carroll, 1993).

Based on these assumptions, the NIH Toolbox CB provides three summary scores: (i) Fluid Cognition Composite, including all the NIH Toolbox CB tests that are fluid ability measures (cognitive flexibility, inhibitory control/attention, episodic memory, and processing speed). The Fluid Cognition Composite score is derived by averaging the standard scores of each of the measures and then deriving standard scores based on this new distribution, (ii) Crystallised Cognition Composite, derived by averaging the vocabulary and the reading test and then deriving standard scores based on this new distribution, and (iii) Global Cognition Function Composite, derived by averaging the Fluid and Crystallised standard scores and then deriving standard scores based on this new distribution. Excellent test–retest reliability, robust developmental effects across childhood, and strong correlations with established, gold-standard measures of similar abilities are suggested (Akshoomoff et al., 2013). However, applying these scores based on averaging of measures presumes that the same abilities are being measured in the same way at different ages (Mungas et al., 2013).

The advantage of these averaged composites lies in creating composites comparable across the lifespan. Furthermore, complex multidimensional constructs can be summarised which facilitates interpretation compared to looking at many indicators separately. Another advantage of the composite approach is the potential to determine developmental trajectories of cognitive development by comparing complex constructs across time (Joint Research Centre-European Commission, 2008). However, for certain research questions, the development of cognitive composite scores may be preferred using statistical approaches such as factor analytic methods. While the advantage is to be able to derive scores that reflect the actual association with different cognitive measures, the disadvantage is that different compositions of factors with different measures and tasks at different age ranges may be limited in their application for longitudinal analyses (Heaton et al., 2014; Weintraub et al., 2013).

There is a range of research on the factor structure underlying cognitive test batteries, with most studies focusing on the dimensions inherent in the Wechsler Adult Intelligence Scale and Wechsler Memory Scale. The frequently identified factors represent more narrowly defined abilities such as verbal abilities, visual perceptual abilities, attention/concentration, and memory (Bowden, Carstairs, & Shores, 1999; Smith et al., 1992).

An earlier study (Mungas et al., 2013) used NIH Toolbox CB data as well as data from several gold-standard measures administered in children and adolescents. They conducted a confirmatory factor analysis (CFA) including the NIH Toolbox CB and gold-standard data to test several *a priori* defined dimensional models. These included a one-factor model representing a global cognition model, a two-factor model representing crystallised and fluid abilities, and several models with different levels of differentiation of fluid and crystallised abilities. In children aged 3–6 years, a three-factor structure was revealed: Vocabulary, Reading, and Fluid abilities. Reading and vocabulary abilities were well distinguished from each other and other cognitive abilities while fluid abilities were not so well differentiated in this age group. The same study identified a five-factor structure (Vocabulary, Reading, Episodic Memory, Working Memory, and Executive Function/Processing Speed) in a separate sample of 8–15-year-olds (Mungas et al., 2013). A study with the NIH Toolbox CB measures as well as measures from several gold-standard tests administered to an adult sample (20–85 years of age) also found five dimensions to explain associations amongst NIH Toolbox CB and gold-standard tests: Vocabulary, Reading, Episodic Memory, Working Memory, and Executive Function/Processing Speed (Mungas et al., 2014).

In summary, there are both fine-grained and broad characterisations of cognition provided by the NIH Toolbox CB (Mungas et al., 2014). An overarching goal of the NIH Toolbox CB is to be administered in large-scale epidemiologic studies and to allow cross-study comparisons. However, earlier studies regarding the factor structure of the NIH comprised of small sample sizes and covered broad age ranges. Furthermore, the applicability of composite measures requires the confirmation of the proposed structures within specific age ranges and to explore alternative factor models and their robustness within sociodemographic diverse cohorts.

It is also important to account for measurement invariance at different ages, across gender, different ethnic groups, and sociodemographic profiles when exploring the factor structure of the NIH Toolbox CB. To date, this has not been examined within the Aotearoa New Zealand context. The large birth cohort study *Growing Up in New Zealand* provides an ideal opportunity to investigate this. The aim of the current study is to:

(i) investigate whether the factor structure of a Global Cognition score, as well as a Fluid and Crystallised Cognitive composite, can be confirmed at the age of 8 years and
(ii) explore what factor structure fits best in a large ethnically diverse cohort of 8-year-old children in Aotearoa New Zealand.

METHOD

Participants

Growing Up in New Zealand is a prospective cohort study with 6822 pregnant women recruited *via* 3 contiguous District Health Board regions in Aotearoa New Zealand

https://doi.org/10.1017/S1355617720001265 Published online by Cambridge University Press

who had expected delivery dates between 25 April 2009 and 2 March 2010 (Morton et al., 2014). The study's cohort is broadly generalisable to current births statistics in Aotearoa New Zealand regarding ethnicity, maternal age, and socioeconomic status (SES) (Morton et al., 2015). A detailed description of the study's design and recruitment can be found elsewhere (Morton et al., 2014, 2015). In brief, major data collection waves (DCWs) have included conducting computer-assisted telephone and personal interviews to gather information longitudinally relating to six interconnected domains of child development: health and well-being, cognitive and psychosocial, education, family and whānau (extended family), culture and identity, and neighbourhoods and societal context.

For the current study, children were included if complete information was obtained from child cognitive observations at the 8-year DCW and if the information on the sociodemographic variables of interest (gender, mother's ethnicity, and SES) was provided ($n = 4298$).

Measures

NIH Toolbox® for Assessment of Neurological and Behavioral Function Cognition Battery (NIH-CB): To assess child cognitive functioning, at the age of 8 years, the NIH Toolbox CB was administered. The NIH Toolbox CB has been validated against existing gold-standard measures, normed in both English and Spanish languages with a sample of 4859 participants (age ranges 3–85 years), has demonstrated strong psychometric properties across the paediatric age range, and tested in typically developing children (Weintraub et al., 2013). The NIH Toolbox CB data were collected using standardised administration procedures and trained interviewers administered the tests. The version 7–17 years was applied to our cohort using the NIH Toolbox iPad app. It comprised the following seven subtests:

Picture Vocabulary Test (PVT): The PVT is a measure of language, particularly assessing receptive vocabulary. The test is administered in a computerised adaptive format where the participant is presented with an auditory recording of a word, paired with four images on the computer screen in which they indicate the image that most likely represents the meaning of the word (Akshoomoff et al., 2013). Item Response Theory (IRT) is used to score the PVT. A theta score (similar to a Z score with $M = 0$ and $SD = 1$) is calculated for each participant. Strong test–retest reliability of the PVT has been found with intraclass correlation (ICC) of .80 (95% CI: .71, .86), as well as high convergent and discriminant validity (Gershon et al., 2014).

Flanker Inhibitory Control and Attention Test: The Flanker Inhibitory Control and Attention Test measures inhibitory control and visual attention. This task is a version of the Eriksen flanker task, derived from the Attention Network Test (Rueda et al., 2004). On each trial (20 items in total), participants are to indicate the left–right orientation of a centrally presented target. Congruent trials represent flankers facing the same direction as the target; on incongruent trials, they face the opposite direction. A computed score is based on a combination of accuracy and reaction time (range 0–10). The Flanker Inhibitory Control and Attention Test showed excellent test–retest reliability (ICC=.92, 95% CI: .86, .95) and adequate to good convergent and discriminant validity amongst 8–15-year-olds (Zelazo et al., 2013).

Pattern Comparison Processing Speed Test: The test measures the speed of choice and response to stimuli and is modelled after Salthouse's Pattern Comparison Task (Salthouse & Meinz, 1995). In this task, the screen shows two side-by-side visual patterns one pair at a time. The participant is given a total of 85 s to respond if the two patterns are the same or not and to do this with as many items as possible (maximum of 130). The participant's raw score is the number of items answered correctly in 85 s of response time (range 0–130). The Pattern Comparison Processing Speed Test demonstrated strong test–retest reliability (ICC=.84, 95% CI: .75, .90) and satisfactory convergent validity (Carlozzi, Tulsky, Kail, & Beaumont, 2013).

List Sorting Working Memory Test (LSWM): LSWM evaluates children's ability to recall and work with recent information. This task is adapted from Mungas' List Sorting task (Mungas et al., 2005) and evaluates the size order sequencing of a familiar stimuli. A series of illustrated pictures are presented with an audio recording of the object's name (One-List and Two-List conditions). In the One-List condition, children recall a series of objects (same category) in the size order of smallest to largest. In the Two-List condition, stimuli are derived from two different categories, in which children recall the objects from one category first in the same size order before recalling the objects from another category in the same size order. The test ends when a participant does not succeed at a given level of difficulty. LSWM is scored by summing the total number of items correctly recalled and sequenced on both lists (range 0–26). For ages 3–15 years, the LSWM showed good test–retest reliability (ICC=.86, 95% CI: .78, .91) and adequate convergent and discriminant validity (Tulsky et al., 2013).

Dimensional Change Card Sort Test (DCCS): The DCCS is a measure of cognitive flexibility and tests an individual's ability to switch focus during a task. The test involves matching a target visual stimulus to one of two choice stimuli according to shape or colour (Akshoomoff et al., 2013). A stimulus image (one of two shapes in one of two colours) is presented on a screen alongside two response images *SHAPE* or *COLOUR*. The participant is instructed to choose the response image that matches the stimulus image in shape or colour (30 items). The scoring procedure is the same as for the Flanker Inhibitory Control and Attention Task (see above). The DCCS has shown strong test–retest reliability (ICC = .92, 95% CI: .86, .95) and adequate to good convergent and discriminant validity (Zelazo et al., 2013).

Picture Sequence Memory Test (PSMT): The PSMT evaluates episodic memory and tests an individual's ability to reproduce a specific sequence of stimuli. Pictures are presented at the centre of the screen and are then moved,

https://doi.org/10.1017/S1355617720001265 Published online by Cambridge University Press

one at a time, into a fixed spatial order. Subsequently, the pictures return to the centre of the screen in a random order and the participant is instructed to recall and reproduce the pictures by placing them in the correct order on the touch screen (Weintraub et al., 2013). Two series with a sequence of 15 and 18 pictures were given. The PSMT is scored using the IRT methodology. The number of adjacent pairs placed correctly in each of the trials is converted to a theta score. The PSMT showed good test–retest reliability for children aged 3–15 years (ICC=.76, 95% CI=.64–.85) and adequate convergent validity (Bauer et al., 2013).

Oral Reading Recognition Test (ORR): The ORR is a measure of reading, evaluating the children's ability to pronounce single words and recognise letters (Weintraub et al., 2013). A series of words and letters are presented on the screen one at a time, which the children are asked to read aloud (Akshoomoff et al., 2013). The test is computer-adaptive, with items being presented in order of increasing difficulty. IRT was used to score the ORR with a theta score being calculated for each participant. The ORR showed excellent test–retest reliability (ICC = .90, 95% CI: .85, .93) as well as convergent and discriminant validity for ages 3–15 years (Gershon et al., 2014).

Additionally, the NIH Toolbox CB provides age-corrected standard scores for each test. This score compares the score of the test-taker to those in the NIH Toolbox nationally representative normative sample at the same age. The normative data in the app follow the US 2010 Census (Casaletto et al., 2015). Both, raw scores and age-adjusted normative scores were used to see if results are comparable within the Aotearoa New Zealand context.

Sociodemographic variables. Information on mother's ethnicity was collected by self-report in the antenatal questionnaire, asking what ethnic groups(s) they belong to and what their main (self-prioritised) ethnicity is. Participants were able to provide a description of their ethnic group at a detailed level, referring to a list of 32 possible answers as well as an open-ended 'Other, please specify' category (multiple responses were collected). In the current study, self-prioritised ethnicity was utilized and categorised into five categories (Statistics New Zealand, 2005): European, Māori, Pacific Peoples, Asian, and Other (including Middle Eastern, Latin American, and African). External prioritisation as per the Statistics New Zealand prioritisation guidelines was used for a small group of mothers with multiple ethnicities who did not choose one main prioritised ethnicity (Statistics New Zealand, 2004).

To determine SES, the NZDep2013 Index was used, which is an area-level measure using socio-economic indicators from the 2013 NZ census (Atkinson, Salmond, & Crampton, 2014). The deprivation categories are area-based measures of deprivation and are assigned to households according to where they live. Deprivation areas are divided into deciles for the whole population and range from least deprived (decile 1) to most deprived (decile 10). In the current study, SES was categorised into high (deciles 8–10), medium (deciles 4–7), and low (deciles 1–3) deprivation.

The distribution of the sociodemographic characteristics of the sample and the NIH Toolbox CB performance (age-adjusted normative scores) can be found in Supplemental Table 1.

Data Analysis

CFA. CFA was first applied to the data to test whether a model with a Global Cognition score, as well as a model with a Fluid and a Crystallised Cognition score, as proposed by the NIH Toolbox CB, fit our data. Indices used to evaluate the fit of the CFA models included the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standardised root mean square residual (SRMR). CFI values of .95 or greater, RMSEA values less than .06, and SRMR less than .08 were argued to indicate acceptable model fit (Hu & Bentler, 1999). Due to its sensitivity to sample size, chi-square statistics were reported but not used to test model fit (Cheung & Rensvold, 2002). McDonald's omega ω (McDonald, 1999) was calculated as a measure of general factor saturation and reliability. This coefficient estimates the extent that a latent construct represents the common variance of all items and can be interpreted according to Cronbach's alpha (Schweizer, 2011).

Exploratory factor analysis (EFA) and subsequent CFA. If the fit of the CFA was found to be poor, we proposed to use EFA to derive the factor structure that would best fit the data of our cohort. Following the EFA, a subsequent CFA was then carried out to examine if the factor structure found in the EFA can be confirmed. Prior to EFA, a random split of the total sample ($n = 4298$) was conducted, generating two subsamples of $n = 2149$ each for EFA and CFA, respectively. We used EFA with maximum likelihood estimation and oblique rotation. In order to identify the number of factors to retain, it is recommended to obtain information from multiple sources (Hayton, Allen, & Scarpello, 2004). We used the Scree test (Cattell, 1966) and parallel analysis (Horn, 1965) as well as theoretical considerations. After identifying the number of factors, we assessed whether the item loadings followed a conceptual logic. Factor loadings of >.3 were considered as substantial. We then employed ω to test the reliability of each factor. Next, we performed CFA with fit indices, as described above.

Measurement invariance (configural, metric, and scalar invariance) of the best-fitting model was investigated across child's gender, SES, and mother's self-prioritised ethnicity. To test for configural invariance and thus for a qualitatively invariant measurement pattern of latent factors across groups, the best-fitting model identified was tested, with factor loadings and thresholds free to vary across groups. Metric invariance is assumed if the item loadings on each factor are invariant across groups, which implies that groups are responding to items in the same way. Scalar invariance is conditioned on metric invariance and additionally assumes item intercepts to be invariant between groups, ensuring that group differences in the means of items are due to differences in the

https://doi.org/10.1017/S1355617720001265 Published online by Cambridge University Press

means of the underlying constructs (Lee, 2018; Stone et al., 2013). Testing for invariance was based on the change in different model fit criteria. Metric invariance can be assumed when the change between models is .01 or less for CFI, .015 or less in RMSEA, .03 or less in SRMR, and .01 or less for scalar invariance (Chen, 2007). We excluded the 'Other' ethnic group from the measurement invariance analyses (across ethnicity) because of its particular heterogeneous composition and small group size ($n = 139$).

Since there is no normative data of the NIH Toolbox available for the Aotearoa New Zealand population and to ensure applicability, all analyses were carried out using the raw scores (PVT, PSMT, and ORR theta score; computed scores for Flanker Inhibitory Control and Attention Test, Pattern Comparison Processing Speed Test, LSWM, and DCCS) as well as using the age-adjusted normative scores of each measure. We used age-corrected scores rather than uncorrected norms as these standard scores were derived separately for children (aged 3–17 years) and adults.

Of 5012 children interviewed at the 8-year DCW of *Growing Up in New Zealand*, 4420 children had complete data in the NIH Toolbox CB observation (note, further 125 cases had to be excluded due to missing sociodemographic data). We compared cases with missing data to complete cases regarding their sociodemographic profile. To check for any bias, we conducted multiple imputation and performed the CFA analysis analogously to the complete cases to provide reassurance if similar results are obtained (White & Carlin, 2010).

Analyses were carried out using RStudio version 1.1.414 and version 25.0 IBM SPSS Statistics.

Ethics

The *Growing Up in New Zealand* study had ethical approval of the Ministry of Health Northern Y Regional Ethics Committee in New Zealand (NTY/08/06/055) and each DCW has been given approval by the Health and Disability Ethics Committee. All procedures using human subjects were conducted in accordance with the standards of the University of Auckland, the Regional District Health Board, and the Health and Disability Ethics Committee. Written informed consent was obtained from mothers for their own and their children's participation.

RESULTS

CFA of the Global Cognition score and the Fluid and Crystallised Cognition Composites. Factor loadings for raw as well as age-adjusted normative scores are presented in Table 1. For the Global Cognition score (Model 1), all measures loaded substantially onto one factor. Factor reliability was below the acceptable threshold ($\omega < .70$). Model fit was acceptable for SRMR but not considered adequate fit with respect to RSMEA, Tucker–Lewis index (TLI), and CFI (see Table 2). For the Fluid and Crystallised Cognition Composites (Model 2), all measures were loading substantially onto the two factors, respectively (see Table 1). Factor reliability for both factors was below the acceptable threshold ($\omega < .70$). Model fit was acceptable for SRMR but not for RSMEA, TLI, and CFI (see Table 2).

Factor structure exploration and validation. After the Scree test and parallel analysis suggested a two-factor structure with loadings that were not plausible from a theoretical perspective, a final three-factor solution (Model 3) emerged after performing EFA, with similar results for the raw and age-adjusted normative scores. Factor loadings are shown in Table 1. Inhibitory Control/Attention, Cognitive Flexibility, and Processing Speed loaded together on one factor (Fluid Cognition I). Vocabulary and Reading loaded together on a second factor (Crystallised Cognition). Working Memory and Episodic Memory loaded together on a third factor (Fluid Cognition II). Factor reliability was below the acceptable threshold ($\omega < .70$) for all three factors. In contrast to the other models, the three-factor model reached an adequate model fit in all fit indices (see Table 2). Overall, factor analysis results of the raw and age-adjusted scores are comparable with slight variations in the factor loadings and reliability of the factors. Substantial correlations amongst the three factors were found, with the highest correlation between Crystallised Cognition and Fluid Cognition II (raw scores: $r = .94$; age-adjusted normative scores: $r = .91$). The correlation between Fluid Cognition I and Fluid Cognition II was slightly higher (raw scores: $r = .71$; age-adjusted normative scores: $r = .58$) than between Fluid Cognition I and Crystallised Cognition (raw scores: $r = .64$; age-adjusted normative scores: $r = .51$).

Measurement invariance testing. The results of the invariance testing can be found in Table 3. For gender, configural, metric, and scalar invariance was revealed for raw as well as age-adjusted scores. With respect to ethnic groups, invariance testing showed configural but neither metric nor scalar invariance for both raw and age-adjusted normative scores. For SES, configural but neither metric nor scalar invariance was found for raw scores; configural and metric but no scalar invariance was revealed for the age-adjusted normative scores.

As we could not confirm measurement invariance, we performed EFA separately by ethnic group to assess whether similar factor structures would be derived across groups. The results can be found in Supplementary Table 2. The three-factor structure proposed in Model 3 could only be identified for the European group. There was a slightly different factor structure, more factor variability, and cross-loadings for Māori, Asian, and Pacific Peoples.

Compared with complete cases, child participants with missing data in the NIH Toolbox CB observation were more likely to live in a highly deprived area, to be of male gender, their mothers were less likely to be of European ethnicity (chi-square test: $p < .01$). Following multiple imputation to account for missing data, CFA was performed analogously to the complete cases with similar results (data not shown).

https://doi.org/10.1017/S1355617720001265 Published online by Cambridge University Press

Table 1. Standardised factor loadings of the Global Cognition score model, the Fluid and Crystallised Cognition scores model (total sample), and the explored model (sample after the random split)

	Model 1	Model 2		Model 3		
Measure	Global Cognition	Fluid Cognition	Crystallised Cognition	Fluid Cognition I	Crystallised Cognition	Fluid Cognition II
Raw scores						
Flanker Inhibitory Control/Attention Test	.54	.57		.61		
DCCS Cognitive Flexibility	.59	.66		.68		
Pattern Comparison Processing Speed Test	.48	.51		.55		
List Sorting Working Memory Test	.50	.47				.69
Picture Sequence Episodic Memory Test	.37	.36				.37
Picture Vocabulary Test	.43		.55		.55	
Oral Reading Recognition Test	.47		.62		.62	
Ω	*.51*	*.44*	*.51*	*.39*	*.51*	*.51*
Age-adjusted normative scores						
Flanker Inhibitory Control/Attention Test	.56	.60		.61		
DCCS Cognitive Flexibility	.62	.64		.68		
Pattern Comparison Processing Speed Test	.52	.54		.54		
List Sorting Working Memory Test	.40	.37				.65
Picture Sequence Episodic Memory Test	.34	.33				.37
Picture Vocabulary Test	.32		.52		.50	
Oral Reading Recognition Test	.35		.58		.59	
Ω	*.60*	*.61*	*.46*	*.61*	*.46*	*.42*

Note. Models 1 and 2 comprise the total sample; Model 3 comprises the sample after the random split ($n = 2149$).ω = McDonald's omega (McDonald, 1999). PVT = Picture Vocabulary Test; FLANKER = Flanker Inhibitory Control and Attention Test; LISTSORT = List Sorting Working Memory Test; DCCS = Dimensional Change Card Sort Test; PATCOMP = Pattern Comparison Processing Speed Test; PSM = Picture Sequence Memory Test; ORR = Oral Reading Recognition Test.

Table 2. Fit indices for models of NIH Toolbox Cognitive Battery dimensions

Model	χ^2	df	RMSEA (90% CI)	SRMR	TLI	CFI
Raw scores						
Model 1: Global Cognition	680.453	14	.105 (.10; .11)	.062	.742	.828
Model 2: Fluid and Crystallised Cognition	487.018	13	.092 (.09; .10)	.055	.802	.878
Model 3*: Fluid I, Fluid II and Crystallised	44.682	11	.038 (.03; .05)	.024	.966	.982
Age-adjusted normative scores						
Model 1: Global Cognition	752.721	14	.111 (.10; .12)	.070	.673	.782
Model 2: Fluid and Crystallised Cognition	531.835	13	.096 (.09; .10)	.062	.753	.847
Model 3*: Fluid I, Fluid II and Crystallised	46.367	11	.039 (.03; .05)	.025	.959	.978

Note. *Sample after the random split ($n = 2149$).
χ^2 = chi-square test; RMSEA = root mean square error of approximation; CI = confidence interval; SRMR = standardised root mean square residual; CFI = comparative fit index; TLI = Tucker–Lewis index.

DISCUSSION

In the current study, we investigated whether the factor structure of a Global Cognition score, as well as a model with a Fluid and Crystallised Cognition Composite, can be confirmed in our sample. We further explored the structural dimensions of the NIH Toolbox CB measures that would fit best in a large ethnically diverse cohort of 8-year-old children in Aotearoa New Zealand.

Participants were members of the *Growing Up in New Zealand* cohort, which is broadly representative of all current births in Aotearoa New Zealand on several key sociodemographic characteristics (Morton et al., 2015). Testing a model with a Global Cognition score and a two-factor model with a Crystallised and Fluid score indicated poor model fit and unsatisfactory reliability for both the raw and age-adjusted scores in our sample. These findings correspond to those of Mungas and colleagues (2013) who conducted CFA for several models with NIH Toolbox CB measures alongside a number of cognitive tests.

A three-factor solution was found to fit our data best, for the raw scores as well as the age-adjusted normative scores of the NIH Toolbox CB measures: Fluid Cognition I (Inhibitory Control/Attention, Cognitive Flexibility, and Processing Speed), Crystallised Cognition (Vocabulary and Reading),

https://doi.org/10.1017/S1355617720001265 Published online by Cambridge University Press

Table 3. Fit indices for invariance analyses for the three-factor model across child's gender, mother's ethnicity, and socio-economic status for raw scores (age-adjusted normative scores)

	RMSEA	ΔRMSEA	SRMR	ΔSRMR	CFI	ΔCFI
Gender						
1. Configural invariance	.041 (.038)		.022 (.022)		.980 (.980)	
2. Metric invariance	.038 (.034)	.003 (.004)	.024 (.022)	.002 (.000)	.979 (.981)	.001 (.001)
3. Scalar invariance	.041 (.039)	.003 (.005)	.027 (.026)	.003 (.004)	.972 (.972)	.005 (.009)
Ethnicity						
1. Configural invariance	.039 (.037)		.023 (.022)		.981 (.982)	
2. Metric invariance	.044 (.118)	.005 (.081)	.027 (.101)	.004 (.079)	.970 (.761)	.011 (.221)
3. Scalar invariance	.072 (.069)	.028 (.049)	.044 (.043)	.017 (.058)	.903 (.901)	.067 (.140)
Deprivation						
1. Configural invariance	.049 (.048)		.026 (.027)		.970 (.968)	
2. Metric invariance	.053 (.050)	.004 (.002)	.032 (.030)	.006 (.003)	.958 (.975)	.012 (.007)
3. Scalar invariance	.060 (.060)	.007 (.010)	.038 (.037)	.006 (.007)	.935 (.927)	.018 (.048)

Note. Fit indices for age-adjusted normative scores are in brackets. Δ = fit index of constrained model – fit index of less constrained model. Only the absolute value is given. A change of ≥.010 in CFI, of ≥.015 in RMSEA, of ≥.030 in SRMR would indicate noninvariance; for testing intercept or residual invariance, a change of ≥.010 in CFI, of ≥.015 in RMSEA or a change of ≥.010 in SRMR would indicate noninvariance (Chen, 2007).
χ^2 = chi-square test; CFI = comparative fit index; TLI = Tucker–Lewis index; CI = confidence interval; RMSEA = root mean square error of approximation; SRMR = standardised root mean square residual.

and Fluid Cognition II (Working Memory and Episodic Memory). The three-dimensional model showed excellent model fit; however, reliability for the three factors was below the acceptable threshold. Fluid Cognition I was more highly correlated with Fluid Cognition II than with Crystallised Cognition. A very strong association was found between Crystallised Cognition and Fluid Cognition II demonstrating a high proportion of shared variance. These findings are in line with those of Mungas and colleagues (2013) regarding factor intercorrelations for 8–15-year-olds and are plausible from a neuroscientific perspective. Working memory, as the ability to store information while simultaneously carrying out processing operations, is a well-established predictor of individual variation in reading comprehension performance in children (Oakhill, Cain, & Bryant, 2003) and has also been found to impact later reading performance as readers gain more reading experience (Peng et al., 2018). Similarly to our findings, Mungas and colleagues (2014) identified a cross-loading of Working Memory on Episodic Memory. The strong association between those two components is conceptually plausible as evidence exists that some episodic memory processes are mediated by working memory mechanisms (Van der Linden, Meulemans, Marczewski, & Collette, 2000). Generally, higher correlations can be found between the measures grouping together onto one factor (see Supplemental Table 3).

It should be noted that despite the excellent model fit, the three factors identified show low reliability. Thus, while there is a tendency to group onto three factors, the underlying cognitive constructs may still be more differentiated, and the identified factors are limited in their function as broader cognitive constructs. Earlier findings revealed five differentiable dimensions of cognitive abilities in children aged 8–15 years as well as in an adult sample (Mungas et al., 2013, 2014). In our study, the seven NIH Toolbox measures were used, which allows for a maximum of three factors. However, our results support the concept of a more differentiated representation of cognitive abilities, with a clear tendency to differentiate crystallised from fluid cognitive abilities and to further distinguish fluid cognitive abilities in 8-year-old children.

Previous research has shown that sociodemographic factors can have an impact on neurocognitive test performances (Casaletto et al., 2015). It must be noted that differences in neuropsychological testing according to ethnic identity may be accounted for by a range of factors, including cultural biases of the measures (Haitana, Pitama, & Rucklidge, 2010; Ogden & McFarlane-Nathan, 1997). Other potential reasons may be more distal background factors and multiple causal pathways including structural inequities affecting cognitive test performance, as opposed to any direct result of ethnicity *per se* (Casaletto et al., 2015; Williams & Mohammed, 2013).

Nevertheless, accounting for ethnicity as a proxy for these underlying background factors aids in identifying differences between groups to be further investigated (Casaletto et al., 2015). We accounted for measurement invariance across gender, SES, and ethnicity for the identified three-factor solution. Configural, metric, and scalar invariance for gender was found for the raw as well as the age-adjusted normative scores. Thus, we can assume that the three cognitive dimensions have the same meaning to participants across gender in our cohort. For the raw scores, configural invariance was found across three deprivation levels (low, medium, and high) as well as across the different groups of mothers' self-prioritised ethnicity (European, Māori, Pacific people, and Asian), while metric and scalar invariance could not be confirmed. These findings demonstrate that the three dimensions specified by the NIH Toolbox CB measures can be used for overall analyses that include participants with different levels of SES and ethnicities as per our Aotearoa New Zealand population. However, differences in factor

https://doi.org/10.1017/S1355617720001265 Published online by Cambridge University Press

variances and covariances are attributable to ethnicity- and SES-based differences in the properties of the measures. This indicates that caution is needed if the three dimensions of cognition are used to compare cognitive performance across different groups stratified by ethnicity and SES.

Our study has several limitations. As described above, the three-factor structure identified mainly represents participants of European ethnicity, while the factor structure appeared to be slightly different for the other ethnic groups. Similarly, full scalar invariance could not consistently be demonstrated for the retained three-factor model across ethnicity and SES, which may limit the applicability for further analyses. In this regard, even when accounting for full measurement invariance across demographic factors, literature generally acknowledges that cognitive tests may be culturally biased with respect to their content and administration procedure (Haitana et al., 2010; Ogden & McFarlane-Nathan, 1997). Fluid constructs, like attention/inhibitory control, maybe less culturally biased than language tests. An earlier study investigated issues of cultural bias by comparing the scores of the Peabody PVT (which is similar to the vocabulary test provided by the NIH Toolbox CB), obtained by 46 Māori children from 3 different age groups (5–11 years), with scores from the standardisation sample. The vocabulary test appeared to be suitable for use with Māori, although the authors made several suggestions to adjust the administration and interpretation to minimise the impact of cultural bias (Haitana et al., 2010). Additionally, a proportion of the children in our sample are bilingual, with some having another language as English as their primary language which might have affected the cognitive test results. In this study, we analysed both raw and age-adjusted normative scores, with the raw scores showing less measurement invariance. By using the age-corrected standard scores provided by the NIH Toolbox CB app, our sample was compared against a sample of 8-year-olds in the USA who are likely to be demographically different from the 8-year-olds in Aotearoa New Zealand, which is a uniquely ethnically and culturally diverse country. It is important to note that our results are limited in their generalisability across national and agegroup populations and cannot inform on the measurement of NIH Toolbox CB constructs across the lifespan or in samples of 8-year-olds from other countries and/or ethnic backgrounds.

There are several strengths to this study, particularly the large size of the sample that is allowed for validating the factor structure of the NIH Toolbox CB in 8-year-old children. In addition, an analogous analysis after multiple imputations demonstrated similar results, suggesting that our findings are applicable to the general population represented by our cohort. The large sample size would have not only increased statistical power but also ensured that group numbers were large enough to test for measurement invariance of the model across gender, ethnicity, and SES. Earlier factor analytic studies with the NIH Toolbox CB measures were conducted with smaller samples and with broader age ranges. For an instrument such as the NIH Toolbox CB to be administered in a particular age group, it needs to show adequate psychometric properties within a specific age group. Furthermore, to our knowledge, this is the first time the NIH Toolbox CB has been validated in a large multi-ethnic cohort in the Aotearoa New Zealand context.

In summary, the lack of reliability of the factors and limited measurement invariance found in our study points to a limited applicability of the NIH Toolbox CB composite scores across the lifespan. Thus, we recommend using the NIH Toolbox CB measures individually in addition to considering the cognitive composites scores, in particular, when analysing cognitive abilities within a certain age group. We also suggest conducting study-specific factor analyses as the composition of composites may vary across sociodemographic factors. Factor analytic approaches may be preferred to mere averaging of cognitive measures when deriving composite scores. Overall, the findings from the factor analyses were comparable between raw and age-adjusted normative scores (comparing the sample to the US normative sample), with slight variations in factor loadings and reliability scores. However, measurement invariance appeared to be lower when using age-adjusted scores. Thus, we recommend that within the Aotearoa New Zealand context, the most conservative approach would be to use raw scores for analysis. It should be noted that the aim of this study was not to focus on the clinically relevant cases but to draw a general picture of the cognitive dimensions of 8-year-olds in Aotearoa New Zealand. For future directions, it will be valuable to examine how the identified factors map onto different outcomes of child development and to investigate whether early developmental disadvantages affect fluid and crystallised abilities in a different manner.

In conclusion, the current study identified three dimensions of cognitive abilities for the NIH Toolbox CB in a sample of 8-year-old children in Aotearoa New Zealand. Our findings show that at this age, fluid abilities are strongly associated with one another than with crystallised abilities and that fluid ability is to be further differentiated. This dimensional structure allows for comparisons across child's gender, but evaluations across SES and ethnicity must be done with caution. Practical implications are to consider the raw scores of the individual NIH Toolbox CB measures rather than using overall composite scores alone within the Aotearoa New Zealand context.

ACKNOWLEDGEMENTS

Growing Up in New Zealand has been funded by the New Zealand Ministries of Social Development, Health, Education, Justice and the former Pacific Island Affairs (now the Ministry of Pacific Peoples); the former Ministry of Science Innovation and the former Department of Labour (now both part of the Ministry of Business, Innovation and Employment); the former Ministry of Women's Affairs (now the Ministry for Women); the Department of Corrections; the former Families Commission (later known as the Social Policy Evaluation and Research Unit and now disestablished);

Te Puni Kokiri; New Zealand Police; Sport New Zealand; the Housing New Zealand Corporation; and the former Mental Health Commission, the University of Auckland and Auckland UniServices Limited. Other support for the study has been provided by the NZ Health Research Council, Statistics New Zealand, the Office of the Children's Commissioner, and the Office of Ethnic Affairs. The study has been designed and conducted by the *Growing Up in New Zealand* study team, led by the University of Auckland. The authors acknowledge the contributions of the original study investigators: Susan M.B. Morton, Polly E. Atatoa Carr, Cameron C. Grant, Arier C. Lee, Dinusha K. Bandara, Jatender Mohal, Jennifer M. Kinloch, Johanna M. Schmidt, Mary R. Hedges, Vivienne C. Ivory, Te Kani R. Kingi, Renee Liang, Lana M. Perese, Elizabeth Peterson, Jan E. Pryor, Elaine Reese, Elizabeth M. Robinson, Karen E. Waldie, and Clare R. Wall. The views reported in this paper are those of the authors and do not necessarily represent the views of the *Growing Up in New Zealand* Investigators.

FINANCIAL DISCLOSURE

The authors have indicated that they have no financial relationships to disclose that are relevant to this article.

CONFLICT OF INTEREST

The authors declared no potential conflict of interest with respect to the research, authorship, and/or publication of this article.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S1355617720001265

References

Akshoomoff, N., Beaumont, J.L., Bauer, P.J., Dikmen, S.S., Gershon, R.C., Mungas, D., … Heaton, R.K. (2013). VIII. NIH Toolbox Cognition Battery (CB): Composite scores of crystallized, fluid, and overall cognition. *Monographs of the Society for Research in Child Development*, *78*(4), 119–132. doi: 10.1111/mono.12038

Atkinson, J., Salmond, C., & Crampton, P. (2014). *NZDep2013 Index of Deprivation*. Wellington: Department of Public Health, University of Otago.

Bauer, P.J., Dikmen, S.S., Heaton, R.K., Mungas, D., Slotkin, J., & Beaumont, J.L. (2013). III. NIH Toolbox Cognition Battery (CB): Measuring episodic memory. *Monographs of the Society for Research in Child Development*, *78*(4), 34–48.

Bowden, S.C., Carstairs, J.R., & Shores, E.A. (1999). Confirmatory factor analysis of combined Wechsler Adult Intelligence Scale—Revised and Wechsler Memory Scale—Revised scores in a healthy community sample. *Psychological Assessment*, *11*(3), 339.

Carlozzi, N.E., Tulsky, D.S., Kail, R.V., & Beaumont, J.L. (2013). VI. NIH Toolbox Cognition Battery (CB): Measuring processing speed. *Monographs of the Society for Research in Child Development*, *78*(4), 88–102.

Carroll, J.B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge University Press.

Casaletto, K.B., Umlauf, A., Beaumont, J., Gershon, R., Slotkin, J., Akshoomoff, N., & Heaton, R.K. (2015). Demographically corrected normative standards for the English version of the NIH toolbox cognition battery. *Journal of the International Neuropsychological Society*, *21*(5), 378–391.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276. doi: 10.1207/s15327906mbr0102_10

Cattell, R.B. (1971). *Abilities: Their Structure, Growth, and Action*. Boston: Houghton Mifflin.

Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504.

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255.

Commission, J.R.C.-E. (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD Publishing.

Gershon, R.C., Cella, D., Fox, N.A., Havlik, R.J., Hendrie, H.C., & Wagster, M.V. (2010). Assessment of neurological and behavioural function: The NIH Toolbox. *The Lancet Neurology* *9*(2), 138–139.

Gershon, R.C., Cook, K.F., Mungas, D., Manly, J.J., Slotkin, J., Beaumont, J.L., & Weintraub, S. (2014). Language measures of the NIH toolbox cognition battery. *Journal of the International Neuropsychological Society*, *20*(6), 642–651.

Haitana, T., Pitama, S., & Rucklidge, J.J. (2010). Cultural biases in the peabody picture vocabulary test-III: Testing tamariki in a New Zealand sample. *New Zealand Journal of Psychology*, *39*(3), 24–34.

Hayton, J.C., Allen, D.G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*(2), 191–205.

Heaton, R.K., Akshoomoff, N., Tulsky, D., Mungas, D., Weintraub, S., Dikmen, S., … Slotkin, J. (2014). Reliability and validity of composite scores from the NIH Toolbox Cognition Battery in adults. *Journal of the International Neuropsychological Society*, *20*(6), 588–598.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.

Horn, J.L. (1970). Organization of data on life-span development of human abilities. In *Life-Span Developmental Psychology* (pp. 423–466). Elsevier.

Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1–55.

Lee, S.T. (2018). Testing for measurement invariance: Does your measure mean the same thing for different participants? *APS Observer*, *31*(8).

Li, S.-C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., & Baltes, P.B. (2004). Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. *Psychological Science*, *15*(3), 155–163.

McDonald, R.P. (1999). *Test Theory: A Unified Treatment.* Mahwah, NJ: L. Erlbaum Associates.

Morton, S.M., Grant, C.C., Carr, P.E.A., Robinson, E.M., Kinloch, J.M., Fleming, C.J., … Liang, R. (2014). How do you recruit and retain a prebirth cohort? *Lessons Learnt from Growing Up in New Zealand. Evaluation & the Health Professions*, *37*(4), 411–433.

Morton, S.M., Ramke, J., Kinloch, J., Grant, C.C., Carr, P.A., Leeson, H., … Robinson, E. (2015). Growing Up in New Zealand cohort alignment with all New Zealand births. *Australian and New Zealand Journal of Public Health*, *39*(1), 82–87.

Mungas, D., Heaton, R., Tulsky, D., Zelazo, P.D., Slotkin, J., Blitz, D., … Gershon, R. (2014). Factor structure, convergent validity, and discriminant validity of the NIH Toolbox Cognitive Health Battery (NIHTB-CHB) in adults. *Journal of the International Neuropsychological Society*, 20(6), 579–587.

Mungas, D., Reed, B.R., Tomaszewski Farias, S., & DeCarli, C. (2005). Criterion-referenced validity of a neuropsychological test battery: Equivalent performance in elderly Hispanics and non-Hispanic Whites. *Journal of the International Neuropsychological Society*, *11*(5), 620–630. doi: 10.1017/S1355617705050745

Mungas, D., Widaman, K., Zelazo, P.D., Tulsky, D., Heaton, R.K., Slotkin, J., … Gershon, R.C. (2013). VII. NIH Toolbox Cognition Battery (CB): Factor structure for 3 to 15 year olds. *Monographs of the Society for Research in Child Development*, *78*(4), 103–118.

Oakhill, J.V., Cain, K., & Bryant, P.E. (2003). The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes*, *18*(4), 443–468.

Ogden, J.A., & McFarlane-Nathan, G.J.N. (1997). Cultural bias in the neuropsychological assessment of young Maori men. *New Zealand Journal of Psychology*, *26*, 2–12.

Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H.L., … Tao, S. (2018). A meta-analysis on the relation between reading and working memory. *Psychological Bulletin*, *144*(1), 48.

Rueda, M.R., Fan, J., McCandliss, B.D., Halparin, J.D., Gruber, D.B., Lercari, L.P., & Posner, M.I. (2004). Development of attentional networks in childhood. *Neuropsychologia*, *42*(8), 1029–1040. doi: 10.1016/j.neuropsychologia.2003.12.012

Salthouse, T.A. & Meinz, E.J. (1995). Aging, inhibition, working memory, and speed. Journals of Gerontology, 50B(6), 297–306.

Schweizer, K. (2011). On the changing role of Cronbach's α in the evaluation of the quality of a measure. *European Journal of Psychological Assessment*, *27*(3), 143–144.

Smith, G.E., Ivnik, R.J., Malec, J.F., Kokmen, E., Tangalos, E.G., & Kurland, L.T.J. (1992). Mayo's Older Americans Normative Studies (MOANS): Factor structure of a core battery. *Psychological Assessment*, *4*(3), 382.

Spearman, C. (1961). *"General Intelligence" Objectively Determined and Measured*. In J. J. Jenkins & D. G. Paterson (Eds.), *Studies in individual differences: The search for intelligence* (pp. 59–73). Appleton-Century-Crofts. doi: 10.1037/11491-006

Statistics New Zealand. (2004). *Report of the Review of the Measurement of Ethnicity*. Wellington, New Zealand: Statistics New Zealand.

Statistics New Zealand. (2005). Statistical Standard for Ethnicity. Wellington, New Zealand: Statistics New Zealand.

Stone, L.L., Otten, R., Ringlever, L., Hiemstra, M., Engels, R.C., Vermulst, A.A., & Janssens, J.M. (2013). The parent version of the strengths and difficulties questionnaire. European Journal of Psychological Assessment, 29, 44–50.

Tulsky, D.S., Carlozzi, N.E., Chevalier, N., Espy, K.A., Beaumont, J.L., & Mungas, D. (2013). V. NIH toolbox cognition battery (CB): Measuring working memory. *Monographs of the Society for Research in Child Development*, *78*(4), 70–87.

Van der Linden, M., Meulemans, T., Marczewski, P., & Collette, F. (2000). The relationships between episodic memory, working memory, and executive functions: The contribution of the prefrontal cortex. *Psychologica Belgica*, *40*(4), 275–297.

Weintraub, S., Bauer, P.J., Zelazo, P.D., Wallner-Allen, K., Dikmen, S.S., Heaton, R.K., … Carlozzi, N.E. (2013). I. NIH Toolbox Cognition Battery (CB): introduction and pediatric data. *Monographs of the Society for Research in Child Development*, *78*(4), 1–15.

Weintraub, S., Dikmen, S.S., Heaton, R.K., Tulsky, D.S., Zelazo, P.D., Bauer, P.J., … Wallner-Allen, K. (2013). Cognition assessment using the NIH Toolbox. *Neurology*, *80*(11 Supplement 3), S54–S64.

White, I.R. & Carlin, J.B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in medicine*, *29*(28), 2920–2931.

Williams, D.R. & Mohammed, S.A. (2013). Racism and health I: Pathways and scientific evidence. *American Behavioral Scientist*, *57*(8), 1152–1173.

Zelazo, P.D., Anderson, J.E., Richler, J., Wallner-Allen, K., Beaumont, J.L., & Weintraub, S. (2013). II. NIH Toolbox Cognition Battery (CB): Measuring executive function and attention. *Monographs of the Society for Research in Child Development*, *78*(4), 16–33.