Intelligence 79 (2020) 101433

Contents lists available at ScienceDirect

Intelligence

journal homepage: www.elsevier.com/locate/intell

Beyond individual intelligence tests: Application of Cattell-Horn-Carroll Theory

Jacqueline M. Caemmerer[a,*], Timothy Z. Keith[b], Matthew R. Reynolds[c]

[a] Howard University, United States of America
[b] University of Texas at Austin, United States of America
[c] University of Kansas, United States of America

ARTICLE INFO

Keywords:
Cognitive
Cattell-Horn-Carroll (CHC) theory
Cross-battery

ABSTRACT

The purpose of this study was to examine the applicability of Cattell-Horn-Carroll (CHC) theory across six intelligence tests to better understand the cognitive abilities at a broad construct level, as opposed to narrow test level. Nearly 4000 youth aged 6 to 18 were drawn from seven tests' standardization and linking samples and missing data techniques were used to complete cross-battery analyses. Cross-battery confirmatory factor analyses demonstrated support for a CHC model when the Differential Abilities Scale, Second Edition, Kaufman Assessment Battery for Children, Second Edition, Wechsler Intelligence Scale for Children, Third, Fourth, and Fifth Editions, and Woodcock–Johnson III Tests of Cognitive Abilities were analyzed simultaneously. All but one of the 66 subtests mapped on the CHC broad abilities in accordance with prior CHC classifications. Results also indicated overall intelligence (*g*) and fluid reasoning (Gf) were statistically indistinguishable. Findings provide further support that the CHC taxonomy is useful for intelligence test classification, interpretation, and development.

1. Introduction

Individuals' intelligence has been linked to many real-world issues including, but not limited to, academic achievement, years of education completed, occupational performance, income, and health behaviors (Gottfredson & Deary, 2004; Neisser et al., 1996). Intelligence tests are often used in diagnostic psychological assessment across a variety of settings (Jewsbury, Bowden, & Duff, 2016). A number of intelligence tests are available, although each test includes measures that differ in content and response format, and tests have been developed according to different theories. Nevertheless, many tests appear to, and sometimes purport to, measure similar intelligence constructs. Thus, a classification system for intelligence constructs that is applicable across all intelligence tests can facilitate communication regarding the constructs measured, individual intelligence test results, and research on intelligence and intelligence tests (see McGrew, 2009). Such a classification system does exist, but there is limited research that has studied the constructs across test batteries rather than in isolation. Such cross-battery research is essential for validation of that system. The purpose of this study was to fill that void by investigating intelligence constructs measured across six popular individually administered intelligence tests.

2. Cattell-Horn Carroll (CHC) theory

The current study was guided by the work of Raymond Cattell, John Horn, and John Carroll. Because Carroll's three-stratum theory and Horn-Cattell's Gf-Gc theory share many commonalities, the synthesis of the two theories is commonly referred to as Cattell-Horn-Carroll (CHC) theory (McGrew, 1997; Schneider & McGrew, 2018). Factor analytic intelligence research findings have generally supported many aspects of CHC theory, and tests derived from other intelligence theories, and even neuropsychological and executive functioning theories, often conform well to a CHC orientation (Floyd, Bergeron, Hamilton, & Parra, 2010; Jewsbury et al., 2016; Keith & Reynolds, 2010; Salthouse, 2005).

CHC theory posits a three-stratum model of intelligence. A general intelligence factor, *g*, is at the apex of the model in the third stratum. In CHC theory, *g* subsumes 8 to 10 broad abilities at the second stratum. *g* and the broad abilities are interrelated and operate together. The broad abilities include: verbal comprehension/knowledge (Gc), or the breadth and depth of acquired cultural knowledge, including language and information learned inside and outside of school (often referred to as crystallized intelligence); fluid reasoning (Gf), or the ability to solve problems using unfamiliar information or novel procedures that cannot be performed automatically; visual-spatial processing (Gv), the ability

* Corresponding author at: Howard University, School of Education, 2441 4th, Street, NW, Washington, DC 20059, United States of America.
E-mail address: jacqueline.caemmere@howard.edu (J.M. Caemmerer).

https://doi.org/10.1016/j.intell.2020.101433
Received 20 June 2019; Received in revised form 28 January 2020; Accepted 6 February 2020

to use mental imagery to solve problems including mental rotations or pattern identification; short-term or working memory (Gsm/Gwm), the ability to hold information in one's immediate awareness and manipulate it; processing speed (Gs), the ability to perform simple, repetitive tasks quickly; and long-term retrieval (Glr), the ability to store and retrieve information over longer periods of time (Schneider & McGrew, 2018). Recently Schneider and McGrew (2018) have argued that this final ability is better understood as two broad abilities: Gl—learning efficiency of new information in long-term memory and Gr—retrieval fluency of previously learned information from long-term storage. These broad abilities subsume many narrow abilities at the first stratum; these abilities are often indexed by individual intelligence subtests. Although different tests often purport to measure the same CHC broad ability constructs, the subtests within each test battery vary according to task demands, stimuli, and response format. Due to these subtest specific differences, some psychologists question whether these different tests measure the same abilities (see Reynolds, Keith, Flanagan, & Alfonso, 2013).

3. Cross-battery intelligence research

In an attempt to better understand the structure of intelligence tests and the cognitive ability constructs measured across batteries, researchers should analyze data from multiple intelligence tests simultaneously. This type of research expands the application of factor analysis from analyzing individual intelligence tests to joint analyses of scores from multiple tests, referred to as cross-battery confirmatory factor analyses (CB-CFA; Reynolds et al., 2013). CB-CFA is a useful technique for understanding the constructs shared by different measures, and also as a method of evaluating intelligence theory (Keith & Reynolds, 2010). CB-CFA analyses also inform the nature of the broad abilities at the construct level. Most intelligence tests do not include more than two or three measures (often referred to as subtests) of any one broad ability, but analyzing data from multiple batteries at once allows for a much larger number of subtest indicators of each broad ability. This results in more generalizable conclusions about the nature of the cognitive abilities the subtests represent. The same common factors should emerge when the subtest indicator variables are selected from different intelligence batteries (referred to as factorial invariance under selection of variables, Reynolds et al., 2013).

CB-CFA has many benefits, but its application in practice is challenging. CB-CFA requires the administration of multiple intelligence tests to the same participant, requiring excessive time, financial, and participant effort. These demands have generally limited prior CB-CFA studies to the simultaneous analysis of two (Flanagan & McGrew, 1998; Keith, Kranzler, & Flanagan, 2001; Keith & Novak, 1987; Phelps, McGrew, Knopik, & Ford, 2005; Sanders, Mcintosh, Dunham, Rothlisberg, & Finch, 2007; Stone, 1992) or three intelligence tests (Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004; Woodcock, 1990). Nevertheless, prior CB-CFA studies based on CHC theory or Cattell-Horn Gf-Gc theory provide cross-battery quantitative evidence supporting these theories, even though most of the tests, except for the Woodcock Johnson, were not explicitly developed using these theories.

A recent CB-CFA analyzed data from four intelligence tests simultaneously: the Kaufman Assessment Battery for Children, Second Edition (KABC-II), Woodcock-Johnson Tests of Cognitive Abilities, Third Edition (WJ III), and Wechsler Intelligence Scale for Children, Third Edition (WISC-III) and Fourth Edition (WISC-IV; Reynolds et al., 2013). Participants were drawn from the KABC-II concurrent validity sample (n = 423 children aged 6–16). Each child completed the KABC-II and one other test for the purpose of investigating correlations between different test scores. A CB-CFA CHC model fit the data well. Five broad abilities, including verbal comprehension/knowledge (Gc, median standardized factor loading [β] = 0.82), fluid reasoning (Gf, median β = 0.66), visual spatial processing (Gv, median β = 0.65), learning efficiency (Gl, median β = 0.68), and working memory (Gwm, median β = 0.61) were well-measured. Correlations between the broad ability factors ranged from 0.57 (Gv and Gwm) to 0.82 (Gf and Gv). Gf generally had the strongest relation with a second-order g factor (β = 0.98), and could not be reliably distinguished from g. The authors concluded that CHC broad abilities were factorially invariant across the four tests and 44 subtests, providing evidence that CHC theory is applicable across different tests and different tests measure the same CHC constructs similarly.

4. Methodological considerations

Modern missing data methods provide one method to exploit the benefits of CB-CFA to test intelligence theory and examine cognitive abilities at a construct level while overcoming the inherent time and resource demands of CB-CFA. For example, not requiring all participants to complete all of the tests eventually included in the CB-CFA analyses may make a study with more than two or three tests feasible. Missing data analysis techniques, specifically full information maximum likelihood (FIML) estimation, may potentially accommodate the incomplete data. Data with incomplete cases are not discarded as they are in list-wise or pairwise deletion. Rather FIML maximizes data that are available by borrowing information from observed data and not discarding important information provided by variables with missing data; missing data are not imputed or replaced, however. The borrowed information from the observed data improves the power and the accuracy of the estimation process (Enders, 2010; Schafer & Graham, 2002).

Missing data mechanisms have implications for the use of FIML with cross-battery intelligence analyses. Three possible missing data mechanisms explain how missing data relate to the variables under study; Rubin and colleagues created these classifications in 1976 (Enders, 2010). One mechanism, missing completely at random (MCAR), is ideal and means there is no association between scores on the intelligence subtests and the cause of the missingness (Enders, 2010; Graham, Taylor, Olchowski, & Cumsille, 2006). Data are missing for individuals who were randomly missing during the data collection process. Another mechanism, missing at random (MAR), means that missing data are related to other variables in the model but not related to scores on the intelligence subtests if all variables related to the missing data mechanism are controlled. The variables related to the missing data mechanism are not included in the model, but serve as auxiliary variables in the estimation process (Enders, 2010). When FIML is applied to MCAR or MAR data and the model is correctly specified, all parameters estimates are consistent and unbiased (McArdle, 1994; Rubin, 1987). The third mechanism, missing not at random (MNAR), means the cause of missingness is related to scores on the intelligence subtests and parameter estimates may be biased (Enders, 2010).

Some methodologists have suggested that researchers actually plan missingness into the design of a research study prior to data collection in order to increase the number of variables or breadth of data collected in a study (McArdle, 1994). McArdle encouraged the use of such planned missingness designs over 20 years ago and wrote “I want to begin by stating: I like incomplete data and think there should be more of it” (McArdle, 1994, p. 409). For example, a planned missingness research design might have participants complete some tests or items and not others and spread out the incomplete data across participants. Often all participants complete one test, referred to as the linking test, and a subset of other tests. Linking tests are needed because their items are vital to the research questions (Enders, 2010; Graham et al., 2006), but a linking test may not be necessary in all scenarios (Graham et al., 2006).

Missingness in planned missingness designs is intentional and under the control of the researcher (Enders, 2010; McArdle, 1994) and all available data are analyzed simultaneously with FIML or some other analytical technique. It should be possible, however, to apply the same analytical techniques to incomplete data that were not “planned” to be

Table 1
Characteristics and demographic of the seven samples.

Test	KABC2 XBA	WISC4/DAS2	DAS2/WIAT2	KABC2 /KTEA2	KABC2 /WISC5	WISC5/WIAT3	WISC4/WIAT2
Gender (%)							
Male	47.7	50.0	48.2	50.1	48.9	55.2	50.8
Female	52.3	50.0	51.8	49.9	51.1	44.8	49.2
Ethnic/racial background (%)							
White, non-Hispanic	60.6	35.2	60.8	62.2	46.6	50.3	61.8
Hispanic	19.7	27.2	19.2	17.7	35.2	21.0	17.5
African American	10.3	24.8	15.8	14.9	10.2	19.9	15.4
Asian	5.4	6.4	3.7	0.0	2.3	1.7	4.3
Native American	0.9	n.r.	n.r.	0.0	n.r.	n.r.	0.4
Other	1.7	6.4	0.6	5.2	5.7	7.2	0.6
Parents' highest level of education (%)							
8th grade or below	–	4.5	5.4	–	1.1	2.2	5.6
9th - 11th grade	9.4*	12.4	11.3	14.4*	12.5	8.3	11.7
High school diploma or GED	20.0	25.7	25.6	32.5	18.2	24.9	26.7
Some college	34.0	28.7	29.9	30.1	36.4	35.4	31.8
Bachelor's or higher	33.4	28.7	27.9	23.0	31.8	29.3	24.2
Age in years							
Mean age (SD)	11.3(2.5)	11.2(3.5)	11.5(3.6)	10.7(4.0)	11.2(2.9)	11.8(3.1)	11.1(3.2)

Note. n.r. indicates values were not reported in the samples. Asterisks denote samples in which a percentage was reported for 11th grade and below only.

missing within a single study, given the same missing data assumptions can be assumed. For example, a combination of data from different studies with similar purposes (e.g., evaluating construct validity of a test) may be analyzed simultaneously given certain assumptions may be met.

5. Purpose of this study

The primary purpose of this research was to investigate CHC theory across six intelligence tests. The validity of CHC-theory-based broad abilities at a construct, as opposed to a test-specific, level and CHC-based classifications of 66 subtests from different tests were investigated. The study builds on previous research (Reynolds et al., 2013) by incorporating data from two intelligence tests not yet included in CB-CFAs, the Wechsler Intelligence Scale for Children, Fifth Edition (WISC-5) and Differential Abilities Scale, Second Edition (DAS-2). As a result, 22 additional subtest indicators were in included in our expanded cross-battery study and an additional CHC broad ability (processing speed) was studied. The second purpose of this study was methodological: to apply principles of missing data to a cross-battery analysis of seven datasets without a single linking test.

6. Method

6.1. Participants

Participants were 3927 children and adolescents aged 6 to 18 drawn from seven samples. These samples were standardization and linking samples collected by Pearson Assessments during norming and validity studies of their intelligence measures. Sample sizes within each sample ranged from 88 to 2223. Demographic information for the samples is shown in Table 1.

1. The Kaufman Assessment Battery for Children, Second Edition (KABC-II) concurrent validity sample included 347 children (referred to as KABC-II XBA). All children completed the KABC-II and one other test, the Woodcock-Johnson Tests of Cognitive Abilities, Third Edition (WJ III, $n = 89$) and the Wechsler Intelligence Scale for Children, Third Edition (WISC-III, $n = 123$) and Fourth Edition (WISC-IV, $n = 58$). The KABC-II XBA sample was used by Reynolds et al. (2013).
2. The KABC-II and Kaufman Test of Educational Achievement, Second Edition (KTEA-II) sample included 2223 children.
3. The KABC-II and Wechsler Intelligence Scale for Children, Fifth Edition (WISC-V) sample included 88 children.
4. The WISC-V and Wechsler Individual Achievement Test, Third Edition (WIAT-III) sample included 181 children.
5. The WISC-IV and Wechsler Individual Achievement Test, Second Edition (WIAT-II) sample included 532 children.
6. The WISC-IV and Differential Abilities Scale, Second Edition (DAS-II) sample included 202 children.
7. The DAS-II and WIAT-II sample included 370 children.

Participant identification numbers were checked across samples to determine whether the same child participated in multiple standardization or validation samples; 16 duplicates were identified. Duplicate intelligence test entries were combined into one. One child with a duplicate entry completed the KABC-II, WISC-III, and KTEA-II, and the other 15 duplicates completed the WISC-IV, DAS-II, and WIAT-II.

6.2. Study design

Across the different samples, children and adolescents were administered specific sets of tests for validation, standardization, and conorming purposes. In six of the original studies, the purpose was to provide evidence of convergent validity by examining the correlations between scores on one intelligence test and scores on another intelligence test, or to provide evidence of predictive validity for that intelligence test and a standardized achievement test. Data from three achievement tests, the KTEA-II, WIAT-II, and WIAT-III, were related to additional research questions beyond the scope of the current study. KABC-II XBA data resembled a planned missingness design. All participants completed the KABC-II, the linking test, plus another test (in our sample they also completed the WJ III, WISC-III, or WISC-IV). The other six samples did not share the same linking test, but one or both tests in each sample were given in at least one other sample (see Table 2).[1]

[1] It may seem counterintuitive that CFA would be possible, or the results meaningful, given these missing data patterns. One way of understanding the methodology is as a multi-group CFA with constraints across groups to allow estimation, and this is a common method for explaining the reference variable approach (e.g., Keith & Reynolds, 2010; McArdle, 1994). Measured variables for tests not administered to a group are treated as latent variables, with constraints across groups (e.g., factor loadings, error variances) to allow estimation in the group with missing data. See the citations for more detail about the method.

Table 2
Seven samples and how they were linked.

Tests	KABC-II	WJ III	WISC-III	WISC-IV	WISC-V	DAS-II	KTEA-II	WIAT-II	WIAT-III
Samples	347	89	123	58	–	–	–	–	–
KABC-II									
XBA									
KABC-II/KTEA-II	2223	–	–	–	–	–	2223	–	–
WISC-IV/DAS-II	–	–	–	202	–	202	–	–	–
DAS-II/WIAT-II	–	–	–	–	–	370	–	370	–
WISC-IV/WIAT-II	–	–	–	532	–	–	–	532	–
WISC-V/WIAT-III	–	–	–	–	181	–	–	–	181
WISC-V/KABC-II	88	–	–	–	88	–	–	–	–

Note. Values represent the sample sizes for each test.

6.3. Measures

Six intelligence tests and 66 subtests were included in the cross-battery analyses. Age-referenced standardized subtest scores were used for all tests. Previous research has demonstrated the factor invariance for each intelligence measure across the age groups: DAS-II (Keith, Low, Reynolds, Patel, & Ridley, 2010), KABC-II (Reynolds, Keith, Fine, Fisher, & Low, 2007), WISC-III (Keith & Witta, 1997), WISC-IV (Keith, Fine, Taub, Reynolds, & Kranzler, 2006), WISC-V (Reynolds & Keith, 2017a, 2017b), and WJ III (Taub & McGrew, 2004).

6.4. KABC-II

The KABC-II, normed for ages 3 to 18, was developed using CHC and Luria theories (Kaufman, & Kaufman, & N. L., 2004). Sixteen KABC-II subtests are designed to measure five CHC broad abilities: Gf, Gc, Gv, Gwm, and Gl. Age-referenced standardized subtest scores range from 1 to 19, with a mean of 10 and a standard deviation of 3. Average internal consistency estimates ranged from 0.74 to 0.93 in the norming sample. Participants in this study were 6 to 18 years old and drawn from three samples (see Table 2).

6.5. WISC

The Wechsler Intelligence Scales for Children, normed for ages 6 to 16 years 11 months, were not developed using CHC theory. Research with the WISC-IV and WISC-III suggests, however, that the constructs measured in these tests align with CHC theory (Keith et al., 2006; Keith & Witta, 1997). The most recent revision, the WISC-V, is more consistent with CHC theory than previous editions (Reynolds & Keith, 2017a, 2017b). WISC subtests measure five CHC broad abilities: Gf, Gc, Gv, Gwm, and Gs. Age-referenced standardized subtest scores range from 1 to 19, with a mean of 10 and a standard deviation of 3.

Average internal consistency estimates for subtests in the norming samples ranged from 0.80 to 0.96 for the WISC-V (Wechsler, 2014), 0.81 to 0.91 for the WISC-IV (Wechsler, 2003), and from 0.69 to 0.87 for the WISC-III (Wechsler, 1991). In the current study 16 WISC-V, 10 WISC-IV, and 12 WISC-III subtests were analyzed. Participants were 6 through 16 years old and were drawn from five samples (see Table 2).

6.6. WJ III

The WJ III is appropriate for a wide age range, from ages 2 to 90 years or older. The WJ III was developed using CHC theory and is the most comprehensive measure of the range of CHC broad abilities in this study. Age-referenced standardized subtest scores are on a standard intelligence scale with a mean of 100 and standard deviation of 15. Median internal reliability estimates for these subtests ranged from 0.74 to 0.94 in the norming sample (Woodcock, McGrew, & Mather, 2001). Eleven subtests representing six CHC broad abilities, Gc, Gf, Gv, Gwm, Gs, and Gl, were analyzed in the current study. Participants in this study were 7 to 16 years old and drawn from one sample, the KABC-II XBA sample.

6.7. DAS-II

The development of the DAS-II was guided by multiple theoretical orientations, including CHC theory. The DAS-II is appropriate for ages 2 to 17. DAS-II subtests measure six broad abilities: Gc, Gf, Gv, Gwm, Gs, and Gl. Age-referenced standardized subtest scores are *T*-scores with a mean of 50 and standard deviation of 10. Average internal consistency estimates ranged from 0.68 to 0.97 in the norming sample (Elliot, 2007). Fourteen subtests were analyzed and participants in this study were 5 to 17 years old, drawn from two samples (see Table 2).

6.8. Data analyses

Three statistical programs were used to conduct the SEM analyses. IBM Statistical Package for the Social Sciences (SPSS, version 25, 2017) was used to select variables and participants and check the data. Following data preparation, invariance was tested, and model implied correlations were estimated via SPSS Amos, Version 23.0 (Arbuckle, 2015). Then, Mplus (Muthén & Muthén, 2018), version 8 was used to analyze the CB-CFA models.

6.9. Missing data

Data analyzed in the current study were not intentionally collected for the purposes of a planned missingness design or CB-CFA (refer to Reynolds et al., 2013 which analyzed data that were also not collected with that intention). Instead data from seven samples were combined and missing data across the samples were due to the methodological approach of combining data. That is, data were missing in the total sample because participants were intentionally not administered all tests. Thus, missingness was likely not related to scores on the intelligence subtests themselves and data were likely MCAR or MAR. Most of the data sets were collected with the intention of providing data on two tests from each participant. Thus, every participant had incomplete data.

There were 70 missing data patterns in the total sample. Proportionally, 68% of the total sample completed the KABC-II, 30% completed an edition of the WISC, 14% completed the DAS-II, and 2% completed the WJ III. See Table 2 for specific sample sizes per subtest.

SPSS Amos and Mplus handle missing data through the strongly recommended Full Information Maximum Likelihood (FIML) procedure. FIML does not discard cases with incomplete data and instead borrows information from the observed data to maximize power and increase the accuracy of the estimation process (Enders, 2010; Schafer & Graham, 2002). When maximum likelihood estimation is applied to missing completely at random (MCAR) or missing at random (MAR) data all parameter estimates are unbiased in correctly specified models. All variables related to the missing data mechanism must be controlled

for in the analyses of MAR data (McArdle, 1994; Rubin, 1987). Parents' education level, a proxy of socioeconomic status, was controlled for in all analyses via the auxiliary variable command in Mplus to account for a possible relation with missingness.[2] The Mplus auxiliary variable command specifies continuous variables that are not part of the analysis, but are used as missing data correlates in addition to the analysis variables (Muthén & Muthén, 2018). The missing data correlate, parent education, provided the correct number of parameters and the chi-square value test for the analysis models (Asparouhov & Muthén, 2008).

7. Analysis plan

7.1. Sample invariance

Measurement invariance was tested across different samples of youth who completed the same test to determine if the CHC broad ability constructs were measured in the same way across samples. For example, in order to establish the WISC-V broad abilities were measured similarly across the two separate groups of youth who completed the WISC-V (88 youth in the WISC-V/KABC-II sample and 181 youth in the WISC-V/WIAT-III sample) measurement invariance was tested. Three other sample invariance tests included: (1) three WISC-IV samples (WISC-IV/DAS-II, WISC-IV/WIAT-II, WISC-IV/KABC-II XBA), (2) three KABC-II samples (KABC-II XBA, KABC-II/WISC-V, KABC-II/KTEA-II), and (3) two DAS-II samples (DAS-II/WIAT-II and DAS-II/WISC-IV).

Measurement invariance was tested at the first order (broad abilities and subtests) level, and constraints were retained in later steps if they were supported. The first step, configural invariance, tested whether the same factor structure fit the data across groups. Next, weak factorial invariance (also known as metric invariance) tested whether the subtest factor loadings were equal across groups. The third step, strong factorial invariance (also known as intercept invariance), tested whether the subtest intercepts were equal across groups. Finally, strict factorial invariance (also known as residual invariance) tested whether the subtest residual variances were equal across groups (Keith, 2019; Meredith, 1993). Given the missing data patterns it may not be necessary to establish invariance, but testing was completed out of an abundance of caution.

7.2. WISC edition invariance

The three editions of the WISC (WISC-III, -IV, and -V) shared 14 subtests. Although the subtests' names were identical across editions, the item content was different in each edition. Measurement invariance was examined across the three editions to determine if the constructs across editions were equivalent.

7.2.1. Cross-battery first-order CHC measurement model

Due to the cross-battery nature of the analyses, CHC broad abilities that were measured by more than one test battery were modeled. Auditory attention (Ga) was not included because only the WJ III assessed that ability. Therefore, a first-order CFA model with six correlated latent CHC broad factors (Gc, Gf, Gv, Gwm, Gs, and Gl) indicated by 8 to 15 subtests was estimated. Three correlated residual variances were included for the four KABC-II and two DAS-II subtests that included a delayed recall measurement (KABC-II Atlantis and Atlantis Delayed, KABC-II Rebus and Rebus Delayed, and DAS-II Recall of Objects Immediate and Recall of Objects Delayed).

Six subtests were cross-loaded on more than one broad ability factor based on previous CFA results; five of the cross-loadings were tested in the Reynolds and colleagues CB-CFA (2013; one was not tested because the DAS-II was not included in that study). The purpose of testing these cross-loadings in the current study was to investigate what these subtests measure with well-defined factors.

- WISC Arithmetic is a complex task tapping possibly more than two broad abilities, and recent findings suggest Arithmetic may also be a direct measure of *g* (Reynolds & Keith, 2017a, 2017b). WISC Arithmetic is labeled a Gwm measure on the scoring structure of older editions but as a Gf measure on the WISC-V; also, Arithmetic items are timed. In our study WISC Arithmetic is more broadly defined than previous studies because data from the third, fourth, and fifth edition were combined. Cross-loadings of WISC Arithmetic on Gf, Gwm, and Gs were tested.
- WISC Picture Completion is labeled a Gv measure in the scoring structure, but a Gc cross-loading was previously supported (Keith et al., 2006; Reynolds et al., 2013).
- KABC-II Gestalt Closure is labeled a Gv measure in the scoring structure, but a Gc cross-loading was supported (Reynolds et al., 2007; Reynolds et al., 2013).
- KABC-II Hand Movements is labeled a Gwm measure, but a Gf cross-loading was supported (Reynolds et al., 2007).
- WJ III Picture Recognition is a Gv measure in the scoring structure, but may be better represented as a Gl measure (Reynolds et al., 2013).
- DAS-II Verbal Comprehension subtest is a measure of receptive language (Gc), but a Gf cross-loading was previously supported (Keith et al., 2010).

7.2.2. Second-order cross-battery CHC model

After the first-order model was established, *g* was added in a second-order model. A model with a direct path from *g* to WISC Arithmetic was tested and compared to the initial second-order model. Parameters of interest in the final model included the factor loadings of the subtests on their respective latent broad abilities and the factor loadings of the broad abilities on *g*. Also, of interest were omega hierarchical coefficients for the six CHC broad abilities. Model implied correlations between Gc, Gf, Gv, Gl, Gwm, and Gs factors and their corresponding composites were estimated in the second-order model with *g*. Composites were created as phantom formative variables with the paths of all their corresponding subtests fixed to 1 (Gignac, 2007). The squared implied correlation is equivalent to omega hierarchical (McDonald, 1999). Omega hierarchical coefficients are considered as either reliability or validity indexes (Reynolds & Keith, 2017a, 2017b).

7.2.3. Model evaluation

Single models were evaluated according to multiple measures of fit, as suggested by methodologists (Hu & Bentler, 1998, 1999). Root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), the comparative fit index (CFI), and the Tucker-Lewis index (TLI) were used to assess the fit of single models (Keith, 2019). Cut-off values that suggest good fit are RMSEA below 0.05, SRMR below 0.08, and CFI and TLI values above 0.95 (Hu & Bentler, 1999).

Change in CFI was used to compare the fit of different invariance models; a cut-off value of equal to or less than −0.01 supported the null hypothesis of invariance (Cheung & Rensvold, 2002). Partial invariance, which allows a limited number of differences across groups, was tested if necessary (Keith, 2019). Alternative, nested models were compared using the likelihood ratio test. For non-nested competing models, the adjusted Bayes Information Criterion (aBIC) was examined; smaller aBIC values indicate better fitting models (Keith, 2019).

8. Results

8.1. Descriptive statistics

Subtest sample sizes, means, standard deviations, skewness, and kurtosis estimates are presented in Table 3. The means and standard

[2] Analyses completed without the auxiliary variable were almost identical to those with parent education as an auxiliary variable.

Table 3
Descriptive statistics for cognitive tests.

Tests and subtests	N	*M*	*SD*	Skew	Ku
DAS-II					
Copying (Gv)	178	51.23	8.41	−0.32	1.14
Digits Backward (Gwm)	557	49.82	8.56	−0.32	0.52
Digits Forward (Gwm)	557	49.81	9.78	0.13	0.93
Early Number Concepts (Gf)	178	51.10	8.64	0.40	−0.19
Matching Letter Like Forms (Gv)	178	51.71	9.08	−0.29	0.40
Matrices (Gf)	557	50.21	9.19	0.09	−0.04
Naming Vocabulary (Gc)	178	51.35	9.41	0.65	1.32
Pattern Construction (Gv)	557	49.97	8.65	0.69	1.61
Rapid Naming (Gs)	557	50.47	8.97	0.84	1.97
Recall of Designs (Gv)	556	50.25	8.78	0.01	0.04
Recognition of Pictures (Gv)	557	50.26	9.29	0.65	1.11
Recall of Objects-Immediate (Gl)	557	49.06	10.32	−0.08	0.92
Recall of Objects-Delayed (Gl)	557	50.17	9.46	−0.07	0.33
Recall of Sequential Order (Gwm)	557	50.09	9.48	0.03	0.74
Sequential and Quantitative Reasoning (Gf)	556	50.55	9.09	0.60	1.20
Speed of Information Processing (Gs)	557	51.07	9.21	−0.01	0.37
Verbal Comprehension (Gc, Gf cross-loading)	178	50.15	8.96	1.09	2.15
Verbal Similarities (Gc)	557	50.66	8.48	−0.29	1.24
Word Definitions (Gc)	556	50.19	8.84	0.16	1.14
KABC-II					
Atlantis (Gl)	2654	10.02	3.09	−0.18	−0.06
Atlantis Delayed (Gl)	2435	9.93	2.80	−0.34	−0.13
Block Counting (Gv)	2655	9.97	3.00	−0.02	−0.11
Expressive Vocabulary (Gc)	2656	9.85	2.95	−0.03	0.05
Gestalt Closure (Gv, Gc cross-loading)	619	10.00	2.89	0.05	0.23
Hand Movements (Gwm, Gf cross-loading)	2656	10.08	2.88	0.04	0.11
Number Recall (Gwm)	2657	10.24	2.86	−0.06	−0.07
Pattern Reasoning (Gf)	2656	10.19	2.96	−0.09	0.04
Rebus (Gl)	2657	10.14	3.04	−0.16	0.01
Rebus Delayed (Gl)	2407	10.03	2.96	−0.28	−0.27
Riddles (Gc)	2657	10.14	3.04	−0.16	0.01
Rover (Gv)	2652	10.15	3.02	−0.05	−0.02
Story Completion (Gf)	2653	10.10	2.98	0.02	−0.02
Triangles (Gv)	2656	10.00	2.91	−0.08	−0.14
Verbal Knowledge (Gc)	2657	10.01	2.94	0.00	−0.07
Word Order (Gwm)	2657	9.93	2.83	0.11	0.13
WISC (Version)					
Arithmetic (Gwm, Gf, Gs cross-loadings; III – V)	880	10.32	2.77	0.16	−0.33
Block Design (Gv; III – V)	1178	10.18	2.84	0.07	0.05
Cancellation (Gs; IV – V)	998	10.03	3.02	0.05	0.06
Coding (Gs; III – V)	1178	10.06	2.90	0.20	−0.01
Comprehension (Gc; III – V)	1174	10.25	2.89	−0.06	0.19
Digit Span (Gwm; III – V)	1167	10.04	2.86	0.16	−0.01
Figure Weights (Gf; V)	269	9.92	2.69	−0.11	0.08
Information (Gc; III – V)	1124	10.25	2.84	0.06	−0.20
Letter-Number Sequencing (Gwm; III – V)	1050	10.00	2.82	−0.46	0.76
Matrix Reasoning (Gf; IV – V)	1060	10.21	2.86	0.13	−0.23
Object Assembly (Gv; III)	123	10.34	2.95	−0.22	0.52
Picture Arrangement (Gf; III)	123	10.65	3.45	0.20	−0.24
Picture Completion (Gv, Gc cross-loading; III & IV)	324	10.33	3.00	−0.08	0.48
Picture Concepts (Gf; IV – V)	1060	10.28	2.92	−0.24	0.19
Picture Span (Gwm; V)	269	9.75	2.66	0.06	−0.58
Similarities (Gc; III – V)	1179	10.18	2.87	−0.09	−0.07
Symbol Search (Gs; III – V)	1143	10.21	2.93	−0.15	0.74
Visual Puzzles (Gv; V)	268	10.06	2.62	0.00	−0.53
Vocabulary (Gc; III – V)	1178	10.14	2.91	−0.16	0.08
WJ III					
Analysis-Synthesis (Gf)	87	102.91	17.19	−0.29	0.40
Auditory Working Memory (Gwm)	88	105.40	13.91	0.34	−0.11
Concept Formation (Gf)	89	105.36	13.90	−0.08	0.54
Decision Speed (Gs)	88	100.56	16.18	−0.71	3.75
General Information (Gc)	89	98.37	16.16	−0.31	0.32
Numbers Reversed (Gwm)	89	100.62	14.31	−0.04	0.52
Picture Recognition (Gv, Gl cross-loading)	89	100.79	12.58	0.04	2.84
Spatial Relations (Gv)	89	100.62	11.33	−0.70	1.33
Verbal Comprehension (Gc)	89	102.55	14.24	−0.69	0.56
Visual-Auditory Learning (Gl)	89	94.65	19.76	−0.47	1.86
Visual Matching (Gs)	89	95.84	13.37	0.34	0.08

deviations of the subtests are mostly similar to those of their respective norming samples. As evidenced in Table 3, the subtests are normally distributed; skewness and kurtosis values are well below suggested cut-off points for univariate normality (below 2 and 7, respectively; Curran, West, & Finch, 1996). The table also shows the broad abilities assumed to be measured by each test. The covariance matrix is available upon request from the first author.

8.2. Invariance testing

8.2.1. Sample invariance

Strict factorial invariance—configural, intercept, metric, and residual invariance—was supported across three models (WISC-IV[3], WISC-V, and KABC-II[2], see Table 4), but not for the DAS-II (change in CFI = 0.011). The Early Number Concepts subtest was the largest contributor to the lack of strict factorial invariance on the DAS-II. Partial strict invariance was tested by allowing the residual of Early Number Concepts to freely vary across the two samples. This modification resulted in a reduction of the change in CFI (0.009), thus supporting partial strict invariance for the DAS-II samples. The invariance results suggest the CHC broad ability factors were measured equivalently across the multiple samples of youth who completed the WISC-IV, WISC-V, KABC-II, and DAS-II. As noted earlier, given the missing data patterns testing for invariance may not have been necessary, but we did so to bolster our and others' confidence in our findings.

Because invariance was supported across the multiple samples of the four tests, subtest data from each sample were merged into a single data column for that respective test. For example, DAS-II Matrices subtest scores from the DAS-II/WIAT-II and DAS-II/WISC-IV samples were combined into one data column within a combined dataset.

8.2.2. WISC edition invariance

The merged invariant samples of the WISC-IV and WISC-V from the previous analysis step were used to test edition invariance. Strict factorial invariance was supported across the WISC-III and WISC-IV data and also between the merged WISC-III/-IV data and the WISC-V data (see Table 4). Factorial invariance suggests that although the subtest content varied across the three WISC editions, the CHC broad ability factors were measured similarly across the three editions. Data for the nine subtests that were administered in all three editions (Vocabulary, Similarities, Comprehension, Information, Coding, Symbol Search, Block Design, Arithmetic, and Digit Span), four additional subtests administered in the WISC-IV and WISC-V (Cancellation, Picture Concepts, Matrix Reasoning, Letter-Number Sequencing), and one additional subtest administered in the WISC-III and WISC-IV (Picture Completion) were merged resulting in the merging of data from five samples. These 14 merged subtests are referred to as “WISC” subtests in later analyses (Reynolds et al., 2013).

[3] In the WISC-IV invariance model Gv and Gf were combined into one factor because there was only one measure of Gv, Block Design, and the model would have been under-identified otherwise. Picture Completion is also a measure of Gv, but only one of the three WISC-IV samples included this subtest; therefore we did not include the subtest here. Picture Completion was also completed by participants in the single WISC-III sample (taken from the KABC-II XBA sample), and invariance was tested across the two editions of the tests and is described below.

2 It was not possible to test one KABC-II subtest, Gestalt Closure, from one sample, KABC-II/KTEA-II, for invariance. Gestalt Closure had a significant amount of missing data in that specific sample (n = 193 out of 2223 total participants). Invariance was supported for Gestalt Closure between the two other KABC-II datasets, KABC-II XBA and KABC-II/WISC-V.

Table 4
Invariance testing across samples & editions.

Model name	$\chi^2(df)$	*p*	CFI	Δ CFI	Adj. RMSEA
WISC-V (2 samples)					
Configural invariance	227.85(188)	0.03	0.97	–	0.04
Metric invariance	241.67(199)	0.02	0.97	0.00	0.04
Intercept invariance	249.00(210)	0.03	0.97	0.00	0.04
Residual invariance	263.05(226)	0.05	0.97	0.00	0.04
DAS-II (2 samples)					
Configural invariance	389.89(310)	0.00	0.97	–	0.03
Metric invariance	405.24(324)	0.00	0.97	0.00	0.03
Intercept invariance	415.80(338)	0.00	0.97	0.00	0.03
Residual invariance	467.62(358)	0.00	0.96	0.01	0.03
Partial invariance	459.89(357)	0.00	0.96	0.01	0.03
WISC-IV (3 samples)					
Configural invariance	282.71(157)	0.00	0.97	–	0.05
Metric invariance	307.31(172)	0.00	0.97	0.00	0.05
Intercept invariance	354.51(187)	0.00	0.96	0.01	0.05
Residual invariance	384.30(210)	0.00	0.96	0.00	0.05
KABC-II (3 samples)					
Configural invariance	841.00(280)	0.00	0.97	–	0.05
Metric invariance	865.83(302)	0.00	0.97	0.00	0.04
Intercept invariance	931.72(324)	0.00	0.97	0.00	0.04
Residual invariance	1011.05(356)	0.00	0.97	0.00	0.04
WISC-III & -IV					
Configural invariance	249.02(119)	0.00	0.97	–	0.05
Metric invariance	254.03(125)	0.00	0.97	0.00	0.05
Intercept invariance	277.81(131)	0.00	0.97	0.00	0.05
Residual invariance	319.56(141)	0.00	0.96	−0.01	0.05
WISC-III/IV & -V					
Configural invariance	323.90(169)	0.00	0.97	–	0.04
Metric invariance	334.84(178)	0.00	0.97	0.00	0.04
Intercept invariance	344.745(187)	0.00	0.97	0.00	0.04
Residual invariance	391.22(200)	0.00	0.97	0.01	0.04

8.3. CB-CFA

8.3.1. CB-CFA first-order model

A first-order CB-CFA with six correlated broad ability latent variables was fit to the data across the six intelligence tests. Gc, Gf, and Gv were measured by 15 subtests, Gwm was measured by 12 subtests, and Gl and Gs were measured by 8 subtests based on a priori classifications. Five subtests—KABC-II Gestalt Closure, KABC-II Hand Movements, WISC Picture Completion, DAS-II Verbal Comprehension, and WJ III Picture Recognition—were cross-loaded on two CHC broad ability factors and WISC Arithmetic was cross-loaded on three CHC broad abilities.

The fit of the first-order CHC CB-CFA model was acceptable to well-fitting (χ2 [1320] = 2497.99, CFI = 0.96, TLI = 0.96, RMSEA = 0.02, SRMR = 0.09, aBIC = 334,312.97). The RMSEA, CFI, and TLI values were considered excellent, and the SRMR value was adequate but slightly exceeded the "good" fit threshold. Most of the factor loadings and the three correlated residual variances were statistically significant. Three cross-loadings were not statistically significant— the WJ III Picture Recognition subtest Gv cross-loading (β = 0.19, *SE* = 0.15, *p* = .21; which replicates Reynolds and colleagues CB-CFA finding (2013)), the DAS-II Verbal Comprehension Gf cross-loading (β = 0.25, *SE* = 0.13, *p* = .05), and the WISC Arithmetic Gs cross-loading (β = 0.08, *SE* = 0.04, *p* = .05). Non-significant cross-loadings were retained in all models and were not pruned. Standardized factor loadings ranged from 0.35 (DAS-II Picture Similarities on Gf) to 0.87 (WISC Vocabulary on Gc; cross-loadings not included). In addition, all six broad abilities significantly correlated with each other (see Table 5 for the correlation coefficients). Gv and Gf correlated with each other most strongly. The weakest correlation was between Gc and Gs.

Table 5
Correlations between broad abilities in the First Order Model.

	Gf	Gv	Gc	Gl	Gwm
Gv	0.90				
Gc	0.78	0.67			
Gl	0.75	0.64	0.68		
Gwm	0.65	0.55	0.60	0.54	
Gs	0.57	0.57	0.42	0.59	0.51

Note. All correlations are statistically significant, $p < .01$.

8.3.2. CB-CFA second-order model

The addition of a second-order *g* factor resulted in model fit which was also acceptable to well-fitting (χ^2 [1329] = 2633.99, CFI = 0.96, TLI = 0.95, RMSEA = 0.02, SRMR = 0.09, aBIC = 334,403.09). An alternative model was tested to determine whether WISC Arithmetic is better represented by multiple cross-loadings or as a direct measure of *g*. In this alternative model a direct path from *g* to WISC Arithmetic was estimated and the three cross-loadings on Gf, Gwm, and Gs were removed. The model with a WISC Arithmetic direct *g* loading fit worse than the previous model according to the aBIC and likelihood ratio test (χ^2 [1331] = 2662.00, $\Delta\chi^2$ = 28.01(2), $\Delta p < 0.01$, CFI = 0.95, TLI = 0.95, RMSEA = 0.02, SRMR = 0.09, aBIC = 334,420.90); this alternative model was rejected. Thus, the model with seven cross-loadings was accepted as the final second-order model and was the basis for interpretation.

As shown in Fig. 1, second-order loadings of the broad abilities on *g* were large and statistically significant. Unlike the other broad abilities, Gf's unique variance was not statistically significant from zero, which, along with Gf's very strong factor loading on *g* (β = 0.992), suggests that Gf and *g* were statistically indistinguishable.

Model implied correlations of the six CHC broad ability factors and their corresponding composites were also estimated. Those correlations were also squared to estimate omega hierarchical coefficients (McDonald, 1999). Omega hierarchical coefficients were 0.64 for Gs (*r* = 0.80), 0.69 for Gl (*r* = 0.83), 0.79 for Gwm (*r* = 0.89), 0.85 for Gv (*r* = 0.92), 0.88 for Gf (*r* = 0.94), and 0.92 for Gc (*r* = 0.96). These findings suggest that possible subtest composites for Gc, Gf, Gv, and Gwm would be highly related to the underlying latent variables, and more so than composites for Gl and Gs.

As with the first-order model, all factor loadings were statistically significant with the exception of the same three cross-loadings (Gv WJ III Picture Recognition, Gf DAS-II Verbal Comprehension, and Gs WISC Arithmetic cross-loadings). Gc standardized factor loadings ranged from 0.71 (WISC Comprehension) to 0.87 (WISC Vocabulary), Gf factor loadings ranged from 0.34 (DAS-II Picture Similarities) to 0.74 (DAS-II Sequential and Quantitative Reasoning), Gv factor loadings ranged from 0.44 (DAS-II Recognition of Pictures) to 0.78 (DAS-II Pattern Construction and WISC Block Design), Gs factor loadings ranged from 0.39 (WISC Cancellation) to 0.74 (WISC Symbol Search), Gwm factor loadings ranged from 0.61 (WJ III Auditory Working Memory) to 0.78 (KABC-II Word Order), and Gl factor loadings ranged from 0.43 (DAS-II Recall of Objects Delayed) to 0.80 (KABC-II Rebus Immediate; excluding subtests that were cross-loaded). Overall, these results suggest the subtests from these six tests are all generally good indicators of the six CHC broad abilities. Thus, these cognitive CB-CFA model results suggest that the six CHC broad ability factors were invariant across the six intelligence tests analyzed in this study.

9. Discussion

The purpose of this study was to test the validity of the application of CHC theory to six individually administered intelligence tests (KABC-II, DAS-II, WJ III, WISC-III, -IV, and -V) using cross-battery confirmatory analyses (CB-CFA). The seven datasets were drawn from co-norming,

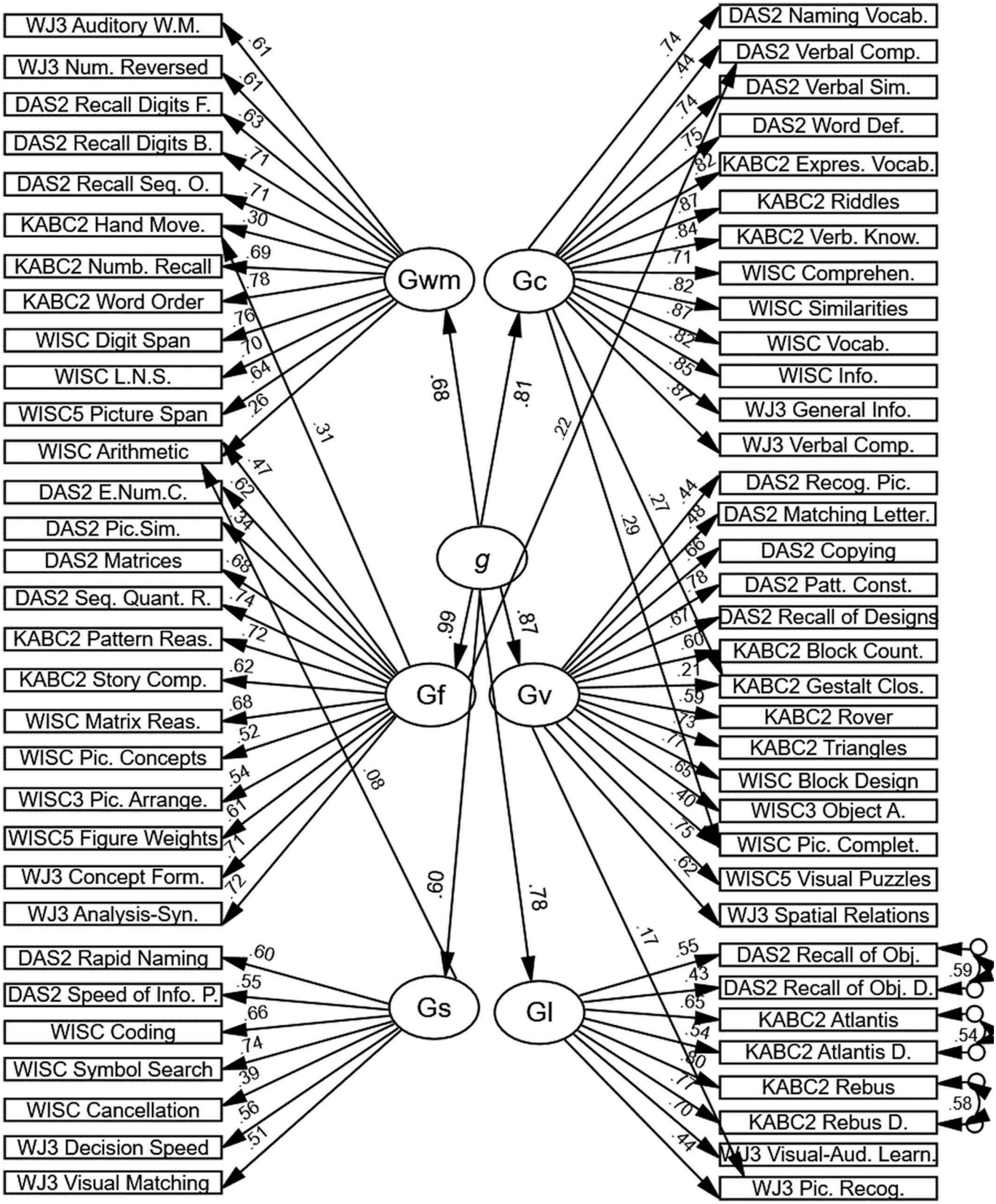

Fig. 1. CB-CFA second-order model standardized coefficients.

standardization, and linking samples, resulting in a large sample covering a broad age range. The analyses used missing data principles and the samples did not share a single linking test. The CB-CFA allowed for an examination of the classifications of 66 subtests and six CHC-theory-based broad abilities were examined at a construct, rather than test-specific level. The results provided further support for the taxonomy of CHC cognitive abilities, regardless of whether or not the tests were explicitly designed using a CHC framework.

9.1. Theoretical implications

A cross-battery CHC model including Gc, Gf, Gv, Gwm, Gl, and Gs fit data from six different intelligence tests well. The factor loadings of all six CHC broad abilities on *g* were large and statistically significant. The pattern of relations and the magnitude of the broad ability factor loadings on *g* is similar to a recent CB-CFA study despite differences in samples and tests included in the two studies (Reynolds et al., 2013). Gf had the strongest loading on *g* and a non-significant residual, supporting previous research that these two constructs are not statistically distinguishable (Caemmerer, Maddocks, Keith, & Reynolds, 2018; Gustafsson, 1984; Reynolds et al., 2013,). Gv had the second strongest loading on *g*, followed by Gc, Gl, Gwm, and finally Gs (β = 0.87, 0.81, 0.78, 0.68, and 0.61 respectively. These findings suggest novel problem solving, visual-spatial problem solving, and the depth and breadth of general knowledge and the ability to retrieve that knowledge efficiently are stronger indicators of overall intelligence, *g*, than are working memory, and simple processing speed. These findings appear to contradict previous assertions that *g* and working memory capacity and *g* and speed of mental operations are equivalent (Colom, Rebello, Palacios, Juan-Espinosa, & Kyllonen, 2004; Jensen, 1993).

At the broad ability level, the Gc, Gf, Gv, and Gwm composites were more highly related to their respective underlying latent variables than the Gl and Gs composites, suggesting higher reliability indexes for these factors. Correlations between all of the latent broad abilities were large and significant. The strongest relation was between Gf and Gv (the same as the 2013 CB-CFA by Reynolds and colleagues), while Gs and Gwm had the weakest correlations with all the broad abilities.

At the subtest level, almost all of the 66 subtests loaded on the broad abilities in accordance with prior CHC classifications (except for the WJ III Picture Recognition subtest which did not load on Gv). A CHC-based interpretation of these 66 subtests from six intelligence tests is well-supported. Subtests with the largest factor loadings may be thought of as the strongest indicators of their respective CHC broad abilities. For example, block construction tasks generally had the strongest loadings on Gv, and rebus tasks had the strongest loadings on Gl, whereas WISC Cancellation and DAS Similarities were weak measures of Gs and Gf, respectively. The four significant cross-loaded subtests (WISC Arithmetic, WISC Picture Completion, KABC-II Hand Movements, KABC-II Gestalt Closure) suggest these subtests may better be conceptualized as mixtures of multiple cognitive abilities (Reynolds et al., 2013). These cross-loaded subtests are supplemental, as opposed to core, subtests and are often not included in composite scores in practice. The three non-significant cross-loadings (Gv WJ III Picture Recognition, Gf DAS-II Verbal Comprehension, and Gs WISC Arithmetic) suggest these subtests are not strong indicators of those broad abilities.

An important theoretical implication based of our results is the perfect correlation between Gf and *g*. Gf and *g* constructs may be redundant and may be used interchangeably. This interchangeable relationship suggests a composite of subtests designed to measure Gf may also be considered primarily a direct measure of *g* (Reynolds et al., 2013). The perfect relation between Gf and *g* raises questions about the structure of intelligence and whether *g* and Gf are redundant or separate abilities (Gustafsson, 1984; Reynolds et al., 2013). The existence of *g* has long been debated in the field. In support of the existence of *g* biological mechanisms of *g*, specifically mitochondrial functioning, have been proposed recently as a form of Spearman's mental energy (Geary, 2018). Another recent theory suggests *g* does not exist because the relations between cognitive abilities are caused by reciprocal interactions during development (mutualism theory, Kan, van der Maas, & Levine, 2019). Still yet another possibility is that *g* is the result of overlapping cognitive processes across diverse tests (Kovacs & Conway, 2019). Further research is needed to better understand the overlap between Gf and *g*.

Taken together, the strong loadings of the broad abilities on *g* and the consistent loadings of the subtests on the broad abilities in accordance with a priori CHC classifications supports the applicability of CHC theory across intelligence tests, regardless of whether the test was explicitly designed with CHC theory in mind. Despite differences across tests in regard to subtest task demands, stimuli, and response formats, these six intelligence tests are measuring the CHC broad abilities similarly, and thus, practitioners and researchers can assume broad ability scores from any of these six intelligence tests are measuring similar underlying cognitive abilities.

The contributions of our CB-CFA CHC model to the cognitive cross-battery literature include the addition of two currently used tests that have yet to be included in such analyses, the DAS-II and WISC-V, 22 additional subtests, the inclusion of an additional broad ability, Gs, and a large sample size. The additions replicate and extend the CB-CFA model presented by Reynolds et al. (2013). This larger CB-CFA CHC model provides further evidence for the applicability of CHC theory to the development of modern intelligence tests, CHC-based classification and interpretation of test results from these six intelligence tests, and cognitive research guided by CHC theory (Reynolds et al., 2013). Use of a common terminology can facilitate communication as CHC theory, and intelligence theory more broadly, continues refinement. Results of this study suggest the validity of CHC terminology beyond applied fields, such as school psychology, to theoretical fields as well.

9.2. Limitations and future research

The findings of this study need to be considered within the context of the study's limitations. This study used principles of missing data analysis, but the data did not include a single linking test. The seven datasets were linked to each other through various configurations of tests they shared in common. While some methodologists have noted a single linking test may not be necessary (Graham et al., 2006) more research is needed to better understand this alternative approach. The influence of the amount of missingness is unknown due to the novelty of the approach, and it is possible that with more observed data between more possible pairs of tests the model fit statistics and other results may be different. Also, the nature of SEM/CFA means there may be other alternative models not tested by the researchers that fit the data as well or better than the model used here.

The current study tested a single model of intelligence against the data from multiple intelligence measures. Future research should expand on the limited theory-testing approach of this study. For example, future research should use cross-battery analyses to test and compare other models of intelligence such as the verbal, perceptual, image rotation (VPR) model supported in previous research (Johnson & Bouchard, 2005). In addition, such data can be used to help evaluate questions about CHC theory, such as the existence and nature of possible intermediate factors, and the validity and consistency of narrow abilities.

Another limitation is findings are limited to the specific tests included in this study and may not be generalizable to other intelligence tests. Due to the specific tests included in this study auditory processing (Ga) was excluded from the analysis because the WJ III was the only test to include measures of Ga. In order to analyze the cross-battery structure of Ga, future research may incorporate other measures in a CB-CFA, such as the newest edition of the WJ, 4th edition, and other types of tests, such as the Comprehensive Test of Phonological Processing, Second Edition (CTOPP-2). Other non-intelligence measures, such as executive functioning tests, can also be used to supplement cognitive tests in future research given findings suggest these neuropsychological measures fit well within the CHC taxonomy (Floyd et al., 2010; Jewsbury et al., 2016; Salthouse, 2005). Such research broadens the use of the CHC nomenclature beyond intelligence tests and further facilitates communication about the abilities various tests share in common.

10. Summary

An adequately fitting cross-battery CHC cognitive model that combines six tests consisting of 66 subtests and seven samples of nearly 4000 youth aged 6 to 18 provides validity evidence for CHC theory. The findings applied to tests and subtests developed from a variety of theoretical orientations, not just those derived from CHC theory. These findings support the applicability of CHC theory to the development and interpretation of modern intelligence tests. Results suggest the CHC classification system is useful even if there are other possible theories that may explain intelligence as well or better. Thus, across applied and theoretical fields CHC terminology can be used as a common language to classify these different cognitive tasks according to overarching broad cognitive abilities.

Acknowledgements

This article is based, in part, on the first author's unpublished doctoral dissertation at the University of Texas at Austin. The authors are grateful to Cindy Carlson and Stephanie Cawthon for their contributions to that dissertation. The authors are grateful to Pearson Assessment and Larry Weiss for data access. Eunice Blemahdoo assisted with manuscript formatting. And we also thank the three anonymous reviewers whose suggestions improved the manuscript.

References

Arbuckle, J. L. (2015). *SPSS Amos 19.0 User's Guide 2015*.

Asparouhov, T., & Muthén, B. (2008). Auxiliary variables predicting missing data. *Technical Appendix*. Los Angeles: Muthén & Muthén.

Caemmerer, J. M., Maddocks, D. L. S., Keith, T. Z., & Reynolds, M. R. (2018). Effects of cognitive abilities on child and youth academic achievement: Evidence from the WISC-V and WIAT-III. *Intelligence, 68*, 6–20. https://doi.org/10.1016/j.intell.2018.02.005.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255. https://doi.org/10.1207/S15328007SEM0902.

Colom, R., Rebello, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. C. (2004). Working memory is (almost) perfectly predicted by *g*. *Intelligence, 32*, 277–296.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16–29.

Elliot, C. (2007). *Differential abilities scale (DAS-II) Manual* (2nd ed.). San Antonio, TX: Harcourt Assessment, Inc.

Enders, C. K. (2010). An introduction to missing data. *Applied Missing Data Analysis*. New York, New York: Guilford.

Flanagan, D. P., & McGrew, K. S. (1998). Interpreting intelligence tests from contemporary Gf-Gc theory. *Journal of School Psychology, 36*(2), 151–182. https://doi.org/10.1016/S0022-4405(98)00003-X.

Floyd, R. G., Bergeron, R., Hamilton, G., & Parra, G. R. (2010). How do executive functions fit with the Cattell–Horn–Carroll model? Some evidence from a joint factor analysis of the Delis–Kaplan executive function system and the Woodcock–Johnson III tests of cognitive abilities. *Psychology in the Schools, 47*, 721–738. https://doi.org/10.1002/pits.

Geary, D. C. (2018). Efficiency of mitochondrial functioning as the fundamental biological mechanism of general intelligence. *Psychological Review, 125*, 1028–1050. https://doi.org/10.1037/rev0000124.

Gignac, G. E. (2007). Working memory and fluid intelligence are both identical to g ?! Reanalyses and critical evaluation. *Psychological Science, 49*(3), 187–207.

Gottfredson, L. S., & Deary, I. J. (2004). Intelligence predicts health and longevity, but why? *Current Directions in Psychological Science, 13*(1), 1–4. https://doi.org/10.1111/j.0963-7214.2004.01301001.x.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11*(4), 323–343. https://doi.org/10.1037/1082-989X.11.4.323.

Gustafsson, J.-E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8*, 179–203.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424–453.

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

Jensen, A. R. (1993). Why is reaction time correlated with psychometric *g*? *Current Directions in Psychological Science, 2*(2), 53–56.

Jewsbury, P. A., Bowden, S. C., & Duff, K. (2016). The Cattell–Horn–Carroll model of cognition for clinical assessment. *Journal of Psychoeducational Assessment, 1*, 1–21. https://doi.org/10.1177/0734282916651360.

Johnson, W., & Bouchard, T. J. (2005). The structure ofhuman intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence, 33*, 431–444.

Johnson, W., Bouchard, T. J., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one g: Consistent results from three test batteries. *Intelligence, 32*, 95–107.

Kan, K.-J., Van Der Maas, H. L. J., & Levine, S. Z. (2019). Extending psychometric network analysis : Empirical evidence against g in favor of mutualism ? *Intelligence, 73*, 52–62. https://doi.org/10.1016/j.intell.2018.12.004.

Kaufman, A. S., & Kaufman, & N. L. (2004). *Kaufman assessment battery for children—Second edition: Technical manual.* Circle Pines, MN: American Guidance Service.

Keith, T., Kranzler, J., & Flanagan, D. (2001). What does the cognitive assessment system (CAS) measure? Joint confirmatory factor analysis of the CAS and the Woodcock-Johnson tests of cognitive ability (3rd edition). *School Psychology Review, 30*(1), 89–119.

Keith, T. Z. (2019). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling* (3rd ed.). New York, NY: Routledge.

Keith, T. Z., Fine, J. G., Taub, G. E., Reynolds, M. R., & Kranzler, J. H. (2006). Higher-order, multi-sample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children—Fourth edition: What does it measure? *School Psychology Review, 35*, 108–127.

Keith, T. Z., Low, J. A., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2010). Higher-order factor structure of the differential ability scales–II: Consistency across ages 4 to 17. *Psychology in the Schools, 47*, 676–697.

Keith, T. Z., & Novak, C. G. (1987). Joint factor structure of the WISC-R and K-ABC for referred school children. *Journal of Psychoeducational Assessment, 4*, 370–386.

Keith, T. Z., & Reynolds, M. R. (2010). Cattell–Horn–Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools, 47*(7), 635–650. https://doi.org/10.1002/pits.

Keith, T. Z., & Witta, L. (1997). Hierarchical and cross-age confirmatory factor analysis of the WISC-III: What does it measure? *School Psychology Quarterly, 12*, 89–107.

Kovacs, K., & Conway, A. R. A. (2019). A unified cognitive/differential approach to human intelligence: Implications for IQ testing. *Journal of Applied Research in Memory and Cognition, 8*, 255–272.

McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research, 29*, 409–454.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.). *Contemporary intellectual assessment* (pp. 151–179). New York: Guildford Press.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*, 1–10.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543.

Muthén, L. K., & Muthén, B. O. (2018). *Mplus user's guide* (Eighth ed.). Los Angeles, CA: Muthén & Muthén.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Ceci, S. J., Loehlin, J. C., & Sternberg, R. J. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77–101.

Phelps, L., McGrew, K. S., Knopik, S. N., & Ford, L. (2005). The general (g), broad, and narrow CHC stratum characteristics of the WJ III and WISC-III tests: A confirmatory cross-battery investigation. *School Psychology Quarterly, 20*(1), 66–88. https://doi.org/10.1521/scpq.20.1.66.64191.

Reynolds, M. R., & Keith, T. Z. (2017a). Multi-group and hierarchical confirmatory factor analysis of the Wechsler Intelligence Scale for Children—Fifth Edition: What does it measure? *Intelligence, 62*, 31–47. https://doi.org/10.1016/j.intell.2017.02.005.

Reynolds, M. R., & Keith, T. Z. (2017b). Multi-group and hierarchical confirmatory factor analysis of the Wechsler Intelligence Scale for Children—Fifth Edition: What does it measure? *Intelligence, 62*, 31–47.

Reynolds, M. R., Keith, T. Z., Fine, J. G., Fisher, M. E., & Low, J. A. (2007). Confirmatory factor structure of the Kaufman Assessment Battery for Children — Second Edition: Consistency with Cattell-Horn-Carroll Theory. *School Psychology Quarterly, 22*(4), 511–539. https://doi.org/10.1037/1045-3830.22.4.511.

Reynolds, M. R., Keith, T. Z., Flanagan, D. P., & Alfonso, V. C. (2013). A cross-battery, reference variable, confirmatory factor analytic investigation of the CHC taxonomy. *Journal of School Psychology, 51*(4), 535–555. https://doi.org/10.1016/j.jsp.2013.02.003.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley.

Salthouse, T. A. (2005). Relations between cognitive abilities and measures of executive functioning. *Neuropsychology, 19*, 532–545. https://doi.org/10.1037/0894-4105.19.4.532.

Sanders, S., Mcintosh, D. E., Dunham, M., Rothlisberg, B. A., & Finch, H. (2007). Joint confirmatory factor analysis of the differential abilities scales and the Woodcock-Johnson tests of cognitive abilities – Third edition. *Psychology in the Schools, 44*(2), 119–138. https://doi.org/10.1002/pits.20211.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147.

Schneider, W. J., & McGrew, K. S. (2018). The Cattell–Horn–Carroll theory of intelligence. In D. P. Flanagan, & E. M. McDonough (Eds.). *Contemporary intellectual assessment* (pp. 73–163). (4thed.). New York, NY: Guilford Press.

Stone, B. J. (1992). Joint factor analysis of the DAS and WISC-R. *Journal of School Psychology, 30*, 185–195.
Taub, G. E., & McGrew, K. S. (2004). A confirmatory factor analysis of Cattell-Horn-Carroll theory and cross-age invariance of the Woodcock- Johnson tests of cognitive abilities ill. *School Psychology Quarterly, 19*, 72–87.
Wechsler, D. (1991). *Wechsler intelligence scale for children, third edition (WISC- III).* San Antonio, TX: The Psychological Corporation.
Wechsler, D. (2003). *Wechsler intelligence scale for children, fourth edition (WISC-IV).* San Antonio, TX: The Psychological Corporation.
Wechsler, D. (2014). *Wechsler intelligence scale for children – Fifth Edition.* Bloomington, MN: Pearson.
Woodcock, R. W. (1990). Theoretical foundations of the WJ-R measures of cognitive ability. *Journal of Psychoeducational Assessment, 8*, 231–258. https://doi.org/10.1177/073428299000800303.
Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III.* Itasca, IL: Riverside Publishing.