Journal of School Psychology 51 (2013) 535–555

Contents lists available at SciVerse ScienceDirect

Journal of School Psychology

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

A cross-battery, reference variable, confirmatory factor analytic investigation of the CHC taxonomy ☆

Matthew R. Reynolds [a,*], Timothy Z. Keith [b], Dawn P. Flanagan [c], Vincent C. Alfonso [d]

[a] University of Kansas, USA
[b] The University of Texas at Austin, USA
[c] St. John's University, USA
[d] Fordham University, USA

ARTICLE INFO

Article history:
Received 4 November 2011
Received in revised form 25 February 2013
Accepted 27 February 2013

Keywords:
Cattell–Horn–Carroll taxonomy
KABC-II
Planned missingness
Fluid intelligence
General intelligence
Flynn effect

ABSTRACT

The Cattell–Horn–Carroll (CHC) taxonomy has been used to classify and describe human cognitive abilities. The ability factors derived from the CHC taxonomy are often assumed to be invariant across multiple populations and intelligence batteries, which is an important assumption for research and assessment. In this study, data from five different test batteries that were collected during separate Kaufman Assessment Battery for Children—Second Edition (KABC-II; Kaufman & Kaufman, 2004) concurrent validity studies were factor-analyzed jointly. Because the KABC-II was administered to everyone in the validity studies, it was used as a reference battery to link the separate test batteries in a "cross-battery" confirmatory factor analysis. Some findings from this analysis were that CHC-based test classifications based on theory and prior research were straightforward and accurate, a first-order Fluid/Novel Reasoning (Gf) factor was equivalent to a second-order *g* factor, and sample heterogeneity related to SES and sex influenced factor loadings. It was also shown that a reference variable approach, used in studies that incorporate planned missingness into data collection, may be used successfully to analyze data from several test batteries and studies. One implication from these findings is that CHC theory should continue to serve as a useful guide that can be used for intelligence research, assessment, and test development.

1. Introduction

Cattell–Horn–Carroll (CHC) theory is an integration of Carroll's three-stratum theory (Carroll, 1993) and the Cattell–Horn extended *Gf-Gc* theory (Horn & Noll, 1997), representing a culmination of over 100 years of psychometric research in human intelligence. It is a multi-factor hierarchical model, which includes 8 to 10 broad cognitive ability factors (e.g., Verbal Comprehension-Knowledge [Gc], Fluid/Novel Reasoning [Gf], Visual–Spatial Ability [Gv], Short-Term Memory [Gsm], Long-Term Retrieval [Glr]), dozens of more specific narrow abilities, and a general factor at the apex (*g*).[1] CHC theory has been used as a template for intelligence test development (Keith & Reynolds, 2010), as a common language used to communicate intelligence test scores and research findings (Flanagan, Alfonso, & Ortiz, 2012; McGrew, 2009), as a model to investigate structural relations between intelligence factors of different order (Kvist & Gustafsson, 2008), and as a model to study relations between intelligence and external variables such as academic achievement (Keith,

☆ We thank Pearson for access to the data used in this research, and the Woodcock-Muñoz Foundation for access to the WJ-III data used to test for measurement invariance.
* Corresponding author at: University of Kansas, Psychology and Research in Education, Joseph R. Pearson Hall, 1122 West Campus Rd, Lawrence KS, 66045-3101, USA. Tel.: +1 785 864 9712.
E-mail address: mreynolds@ku.edu (M.R. Reynolds).
ACTION EDITOR: Craig Albers.
[1] There are differences in opinion on whether a *g* factor should be included, but this discussion is beyond the scope of the current article (cf. Carroll, 2003; Horn & McArdle, 2007).

0022-4405/$ – see front matter
http://dx.doi.org/10.1016/j.jsp.2013.02.003

1999) and personality (Baker & Bichsel, 2006). It is for these reasons, among others, that CHC theory has been recommended as a useful framework for intelligence research (McGrew, 2009). Nevertheless, it is a working model. There are numerous hypotheses that may be tested, including refinement or alterations to the framework itself (e.g., Kan, Kievit, Dolan, & van der Maas, 2011).

Both measurement and structural aspects of CHC theory are investigated in this study, with two primary aims. The first aim is to investigate the invariance of CHC broad ability factors across different intelligence batteries (Mulaik, 2010; Thurstone, 1947). Most new intelligence batteries specify a measurement structure so that two or three subtest scores are grouped to align with CHC broad abilities. If CHC theory is a useful framework for grouping subtest scores, subtests specified to load on CHC factors within a test battery should load on the same CHC factors when subtest scores from other test batteries are included in a factor analysis. The second research aim is to investigate the structural relations between first-order CHC broad ability factors and a second-order *g* factor. Of particular interest are the *g*–Gf relation and the influence of population heterogeneity on these first- and second-order relations (which are sometimes called *loadings*). Gf and *g* are often either perfectly or nearly perfectly related (Gustafsson, 1984), and population heterogeneity may be one reason that loadings vary across studies (Kvist & Gustafsson, 2008).

1.1. Cattell–Horn–Carroll taxonomy

A second-order *g* factor and first-order broad cognitive ability factors are often used to represent an underlying system of common factors that produces variation in intelligence test scores. Almost all tests measure *g*, one or more broad cognitive abilities, and something specific to that test. CHC theory is one such descriptive system (Schneider & McGrew, 2012). It is also often referred to as a taxonomy and has been used extensively to classify abilities and to select, organize, and interpret intelligence test scores (Flanagan et al., 2012). Several intelligence batteries have subtests organized into composites based on CHC theory (e.g., Kaufman Assessment Battery for Children, Second Edition [KABC-II], Kaufman & Kaufman, 2004, and the Woodcock–Johnson III [WJ III] Tests of Cognitive Abilities, Woodcock, McGrew, & Mather, 2001). Each battery assesses various CHC broad abilities, with several abilities (Gf, Gc, and Gv) measured in each. Although CHC factors common to these intelligence batteries are similar in name, they are measured by subtests varying in their task demands, response format, and test stimuli. Are the CHC factors measured by these subtests the same? This question is one of *factorial invariance*.

1.1.1. Factorial invariance

The replicability of common factors across various conditions speaks to the basic nature and usefulness of those factors. Such replicability is encapsulated in factorial invariance. Two types of factorial invariance have been described: invariance under selection of populations and invariance under selection of variables (Mulaik, 2010; Thurstone, 1947). If factors are invariant under selection of populations, then the same common factors should emerge when the same battery is administered to other populations. If factors are invariant under selection of variables, then the same common factors should emerge when other indicator variables (e.g., subtests) are selected from a broader domain of indicators (e.g., different intelligence batteries).

Invariance under selection of populations is often investigated via multi-group confirmatory factor analysis (MG-CFA) and has been formulated within the broader scope of measurement invariance (Meredith, 1993). This approach is often used to investigate internal test bias. Measurement invariance of CHC broad ability factors has been investigated and for the most part supported across age (e.g., Taub & McGrew, 2004) and sex (e.g., Reynolds, Keith, Ridley, & Patel, 2008) in several popular intelligence batteries. The lesser known type of invariance—although implicitly assumed in almost all research and assessment—is invariance under selection of variables (Thurstone, 1947). This type of invariance implies that the factorial composition of a subtest, described via factor loadings, should not differ when a subtest is moved to a new battery, which also includes that common factor. Tucker (1958) succinctly summarized this issue by posing the question "Do factors transcend batteries" (p. 112)? Factors should transcend batteries.

There have been attempts to establish the invariance of *g* under different selections of tests. Thorndike (1987) inserted subtests into different battery groupings, which were then submitted to factor analysis, and found correlations ranging from .52 to .94 between factor loadings of those subtests. He concluded that *g* was invariant because the rank ordering of the loadings was fairly stable across analyses. Other researchers have studied *g* loadings across different test batteries, and have estimated the influences of battery size and composition, factor-extraction technique, and interactions among these influences on the dependability of *g* loadings (cf., Floyd, Shands, Rafael, Bergeron, & McGrew, 2009; Jensen & Weng, 1994; Major, Johnson, & Bouchard, 2011). Despite some similar findings across these studies, the interpretations of those findings have varied.

One argument often levied against *g* invariance is that *g* depends on the characteristics and composition of subtests included in a battery (Horn, 1991). For example, *g* may be biased toward Gc because Gc tests are often overrepresented in intelligence batteries (Ashton & Lee, 2006). Or, the way in which *g* is extracted or modeled, especially related to Gc, may also bias *g*. For example, compared to Gf, "a larger proportion of the systematic variance in the Gc-tests is turned into common factor variance," thus in many studies in which a *g* factor has been represented by a first principal factor, the *g* factor is actually more Gc-like (Kvist & Gustafsson, 2008, p. 434). Because of this concern, among others, higher-order factor models have often been recommended to model *g* (Jensen, 1998; Keith, 2005). When *g* has been modeled as a second-order factor in CFA models using data from examinees who were administered more than one intelligence test battery, the two *g* factors correlate either perfectly or near perfectly (Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004; Keith, Kranzler, & Flanagan, 2001). Some researchers consider this correlation as evidence of *g* invariance. The substantive interpretation of that correlation differs, however. Some researchers suggest that it is strong because the tests measure the same "thing," whereas other researchers suggest it is strong because they measure the same "things" (cf., Horn & McArdle, 2007, p. 221; Johnson et al., 2004; van der Maas, Dolan, Grasman, Wicherts,

Huizenga, & Raijmakers, 2006). Therefore as it stands, *g* factor invariance is somewhat ambiguous, although approaches to a resolution have been offered (Gustafsson, 1984).

Regardless of the disagreements about the nature of or invariance of *g*, a single factor does not adequately account for shared variance among intelligence test scores (Carroll, 1993; Jensen, 1998). Group factors, such as latent CHC broad abilities, are also important and more suitable for invariance tests because they are directly linked with and defined by certain tests. There is research support for invariance of CHC broad ability factors (e.g., Carroll, 1993). Moreover, the CHC taxonomy has been used in recent intelligence test development with each battery developed independently by different research teams and test authors and normed in different samples. Although several of the same CHC factor descriptions are used across batteries, the individual subtests aligned with the factors differ in content, response processes, and surface features. Despite these differences in participant and measurement sampling, an overwhelming majority of subtests load on the expected CHC factor within a battery, even when a battery was not designed to measure CHC factors (Keith & Reynolds, 2010). Nevertheless, this evidence relies mostly on the internal structure of scores within a battery, whereas factorial invariance may be studied with more rigor by investigating the latent structure of subtest scores across intelligence batteries. Factor analysis of subtest scores from multiple batteries administered to the same participants is referred to as joint or cross-battery factor analysis (CB-FA).

1.1.2. CB-FA research

Woodcock (1990) demonstrated the utility of CB-FA for studying theoretical constructs underlying intelligence batteries with his CB-FA of nine data sets that included measures from six different intelligence batteries. The factorial composition of the Woodcock–Johnson Psychoeducational Battery-Revised (WJ-R; Woodcock & Johnson, 1989) subtests was similar whether derived from the WJ-R norming sample or from studies that included subtests from other intelligence batteries, thus providing construct validity evidence for the WJ-R and invariance of the Cattell–Horn extended *Gf-Gc* theory constructs. Since Woodcock's CB-FAs, several similar studies have been conducted. The majority of these studies used Gf-Gc or CHC-driven FAs to evaluate the structural validity of individually administered intelligence batteries (e.g., Flanagan & McGrew, 1998; Keith et al., 2001; Phelps, McGrew, Knopik, & Ford, 2005; Sanders, McIntosh, Dunham, Rothlisberg, & Finch, 2007). Most of these studies, however, used intelligence batteries that have since been revised.

There are a few CB-FAs that have included intelligence measures currently used by school psychologists. For example, Roid (2003) found that a five factor model with Gf, Gc, Quantitative Reasoning, Gv, and Working Memory fit data from a CB-FA with subtests from the Stanford–Binet Intelligence Scales, Fifth Edition (Roid, 2003) and WJ III. Moreover, Floyd, Bergeron, Hamilton, Parra, and McGrew (2010) used scores from the WJ III and Delis–Kaplan Executive Functioning Scales (Delis, Kaplan, & Kramer, 2001) and found that Gc, Gsm, Gv, Glr, and Executive Functioning factors emerged from tests across the two batteries, which were derived from different theories.

Revisions of intelligence batteries that do not include substantial changes to the subtests should not result in "new" factors; nevertheless, updating this research by using scores from current revisions of intelligence batteries provides further evidence of factorial invariance. These findings should be useful for psychologists who use these instruments to make inferences about children's intelligence without full assurance that the abilities measured by one test are the same as those measured in another. Therefore to address the first aim of this study, CB-CFA is used to model subtest scores from current revisions of some of the most popular intelligence batteries (e.g., KABC-II, WJ III, and WISC-IV). There is no known published research using CB-CFA with the WISC-IV or KABC-II in school-age children despite their popularity. Our research thus provides a needed update. Moreover, our study includes data from the WISC-III. If various assumptions are met, the use of these data allows for linkage across Wechsler revisions.

1.2. Second-order g factor loadings

CHC theory may also be used to understand structural relations (a.k.a., factor loadings) between a second-order *g* factor and first-order factors. These relations were investigated in the second aim of this study, with an emphasis on the Gf–*g* relation. In previous within-battery research, first-order factors such as Gsm and Gs have shown relatively lower standardized *g* loadings (in the .50 to .65 range) than Gf, Gc, and Gv factors (e.g., Bickley, Keith, & Wolfle, 1995). Because of the large Gc and Gf loadings on *g*, these two factors have generated the most research interest (Carroll, 1996). The most important and consistent finding is that a Gf first-order factor often demonstrates a near perfect or perfect relation with a second-order *g* factor (Carroll, 2003; Gustafsson, 1984; Undheim & Gustafsson, 1987). In a study using norming sample data from the KABC-II and WJ III, Gf was statistically indistinguishable from *g* within the KABC-II, and other than in one age group, Gf loadings on *g* ranged from .95 to .97 within the WJ III (S. B. Kaufman, Reynolds, Liu, A. S. Kaufman, & McGrew, 2012). Gf was measured by only two or three subtests within each battery, however. Additional second-order loadings across the KABC-II and WJ III batteries in different age groupings of children and adolescents were also reported: Gc loadings ranged from .83 to 90; Gv from .74 to .93; and Glr from .78 to .85. Gsm loadings varied between batteries, ranging from .64 to .71 on the KABC-II, on which Gsm was defined by measures of memory span, and from .84 to 1.00 on the WJ III, on which Gsm was defined more by measures of working memory, which are more complex.

Although there are some consistent findings related to standardized second-order *g* loadings across studies, Arendasy, Hergovich, and Sommer (2008) indicated sample homogeneity and construct representation as two possible reasons for inconsistent findings (cf. Carroll, 1996, 2003). These authors incorporated construct representation into automatic test item generation, utilizing information from cognitive psychology and individual difference research. This approach, long advocated by researchers in test development (Embretson, 1983), was used to automatically generate items for the construction of two

subtests per five CHC broad abilities. Data were collected using these new measures, and a CHC model provided the best description of them in a factor analysis. The Gf–*g* relation approached unity (.98), and the remaining factors had relatively weaker loadings on the *g* factor (e.g., Gsm = .76, Gv = .77, and Gc = .75). Thus, different methods have produced similar findings.

An additional neglected aspect of factor loading interpretations is related to population heterogeneity within a sample (Muthén, 1989). Kvist and Gustafsson (2008) modeled intelligence data with a higher-order CHC model. They found that the Gf–*g* relation reached unity when data were analyzed within each sample of participants who had different learning opportunities. Alternatively, they found that there was a clearly differentiated Gf–*g* relation in the same CHC higher-order model when the samples were pooled for analysis. In addition, the Gc–*g* relation was considerably lower than the Gf–*g* relation within each sample, but it was similar to the Gf–*g* relation when the samples were pooled (Gf–*g* = .83 and Gc–*g* = .80). The findings were interpreted using Cattell's (1987) investment theory, in which "historical Gf" (a.k.a., the second-order *g* factor in that study) underlies shared variation among first-order factors. Moreover, because Gf and *g* were statistically indistinguishable, it was argued that if a first-order Gf factor is identified in an invariant manner, then so too is *g*.

One final, but important limitation of prior Gf–*g* research has been that Gf subtests have been historically underrepresented (Carroll, 2003). In this study, we have multiple measures of Gf, including those measuring different aspects of Gf, and we attempted to control for heterogeneity related to socio-economic status (SES) and sex. We expected that the Gf–*g* loading would approach unity.

1.3. Missing data by design

CB-FA research is valuable for understanding the constructs measured by tests, for testing theory, and for establishing factorial invariance across tests. It is rare, however, in part because data collection is so challenging. CB-FA studies require a time commitment from participants because all subtests from a minimum of two intelligence batteries are often administered. Additional concerns include cost and examinee fatigue. As such, CB-FA studies tend to include no more than two intelligence batteries, relatively limited indicators per factor, and smaller samples.

One efficient method of overcoming these challenges is to use a reference variable method as part of a missing data by design approach to data collection (McArdle, 1994). A *reference-variable approach*, or *planned missingness design*, is not new, but it is rarely discussed outside of the methodological literature (cf. Graham, Taylor, & Cumsille, 2001; Graham, Taylor, Olchowski, & Cumsille, 2006; McArdle, 1994). A 3-form design procedure for incorporating planned missingness with reference variables completed by all participants has been described as an optimal method (Graham et al., 2006). For example, a researcher may be interested in studying one factor with 10 item indicators. Rather than administering all 10 item indicators to everyone, only one or two reference indicators (e.g., items 1 and 6) are administered to everyone, with different subsets of the other indicators administered to participants. Raw incomplete data are then analyzed using maximum likelihood estimation procedures. This approach should be especially useful in designing research across intelligence batteries because every subtest or test battery need not be administered to every participant (see Keith & Reynolds, 2012; McArdle, 1994). A variant of this approach is used in the current study so that subtest scores from multiple test batteries are analyzed simultaneously even though every battery is not administered to every participant.

1.4. Present study

The first aim of this study was to investigate invariance of CHC broad ability factors across different selections of populations and variables. The measures used in this study are current revisions of some of the most popular intelligence batteries, and the findings should inform psychologists who interpret CHC factors within test batteries and who use a cross-battery approach for assessment (Flanagan et al., 2012). Given the broad and diverse nature of the tests included, the study also provides a test of the validity of CHC theory as a model of intelligence across measures. The second aim of this study was to investigate the loadings of first-order factors on a second-order *g* factor, the Gf–*g* relation in particular, and how population heterogeneity may influence these loadings. Findings related to this second aim should inform those who study the psychometric structure of intelligence. Last, although not a primary aim, it is demonstrated how a reference variable approach to a planned missingness design may be used to expand construct representation in a CB-CFA study.

2. Method

2.1. Participants

Participants in this study were 423 children and adolescents, aged 6 to 16 years ($M = 11.08, SD = 2.69$), who were included in the concurrent validity studies for the KABC-II. Individual study samples were intended to represent the United States population. The total sample was representative of the United States population with regard to sex (boys = 209, 49.4%; girls = 214, 50.6%). Race/ethnicity was fairly representative in terms of European American ($n = 267$, 63.1%) and Native American children ($n = 3$, 0.7%); however, Hispanic/Latino ($n = 83$, 19.6%) and Asian American ($n = 20$, 4.7%) children were slightly oversampled, and African American (39, 9.2%) children were slightly undersampled. Race/ethnicity was not reported for 5 children (1.2%) and "other" race/ethnicity was recorded for 6 others (1.4%). Socioeconomic status was indexed by parent education. Children of parents with 11th-grade or less education, who graduated high school or obtained a GED, who had some college, and who earned a bachelor's degree or higher

Table 1
A priori CHC classifications of cognitive ability subtests by battery.

	CHC ability				
Test battery	Gc	Gv	Gf	MA	Gsm
KABC-II	Expressive Vocabulary Verbal Knowledge Riddles *Gestalt Closure*	*Gestalt Closure* Block Counting Rover Triangles	Pattern Reasoning Story Completion *Hand Movements*	Atlantis Atlantis-Delayed Rebus Rebus-Delayed	Word Order Number Recall *Hand Movements*
WISC-III	Information *Picture Completion* Comprehension* Similarities* Vocabulary*	Object Assembly Block Design* *Picture Completion*	Picture Arrangement *Arithmetic*		*Arithmetic* Digit Span*
WISC-IV	Comprehension* Similarities* Vocabulary*	Block Design*	Matrix Reasoning Picture Concepts		Letter–Number Sequencing Digit Span*
WJ III	Verbal Comprehension General Information	Spatial Relations Picture Recognition	Concept Formation Analysis–Synthesis	Visual–Auditory Learning	Numbers Reversed Auditory Working Memory
PIAT-R/NU	General Information				

Note. * = measured in both WISC-III and WISC-IV samples. Italics indicate subtest loaded on two first-order factors initially. Gc = Verbal Comprehension-Knowledge; Gv = Visual Spatial Ability; Gf = Fluid/Novel Reasoning; MA = Associative Memory; Gsm = Short-Term Memory; WISC-III = Wechsler Intelligence Scale for Children—Third Edition (Wechsler, 1991); WISC-IV = Wechsler Intelligence Scale for Children—Fourth Edition (Wechsler, 2003); KABC-II = Kaufman Assessment Battery for Children—Second Edition (Kaufman & Kaufman, 2004); WJ III = Woodcock-Johnson III Tests of Cognitive Abilities (Woodcock et al., 2001); PIAT-R/NU = Peabody Individual Achievement Test-Revised/Normative Update (Markwardt, 1998).

composed 9.0% ($n = 38$), 19.1% ($n = 81$), 35.7% ($n = 151$), and 33.6% ($n = 142$) of the sample, respectively. SES was not reported for 11 children (2.6%).

2.2. Measurement instruments

Subtests from four individually administered intelligence batteries were used in this research: the KABC-II (Kaufman & Kaufman, 2004), the WISC-III (Wechsler, 1991), WISC-IV (Wechsler, 2003), and WJ III (Woodcock et al., 2001). In addition, one subtest from the Peabody Individual Achievement Test-Revised/Normative Update (PIAT-R/NU; Markwardt, 1998) was included. These data were collected as part of concurrent validity studies during the development of the KABC-II. They are from some of the most popular individually administered intelligence batteries currently available and from batteries measure overlapping CHC abilities.

2.2.1. KABC-II

The KABC-II was developed and organized using the CHC taxonomy and Lurian theory (Kaufman & Kaufman, 2004). The battery includes composite scores for four CHC broad abilities and one CHC narrow ability: Gf, Gv, Gc, Gsm, and Associative Memory, respectively. The factor structure of the battery has been supported in research (Reynolds, Keith, Fine, Fisher, & Low, 2007). The KABC-II correlated first-order factor model has been shown to be invariant across age and sex (Reynolds et al., 2007, 2008). Average internal consistency estimates for subtest scores for children aged 7 to 12 years ranged from .74 (Gestalt Closure) to .93 (Rebus) in the norming sample. All 16 standard and supplemental subtests were used in this study. Predicted CHC classifications for these subtests are shown in the top row of Table 1.

2.2.2. WISC-III

The WISC-III (Wechsler, 1991) includes 13 subtests, although only 10 were used in this CB-CFA portion of the study. The Mazes subtest was not administered to participants so it was not included at all, and Gs subtests were not included in the reference battery (i.e., KABC-II), thus the WISC-III Gs subtests were excluded in our CB-CFA. A four-factor model (Verbal Comprehension, Processing Speed, Perceptual Organization, and Freedom from Distractibility) was used to describe the structure of the WISC-III. From a CHC perspective, the Verbal Comprehension and Processing Speed factors are similar to Gc and Gs factors, respectively. The Perceptual Organization factor seems to include tests of Gv (the Object Assembly, Block Design, and Picture Completion subtests) and Gf (the Picture Arrangement subtest[2]). The Freedom from Distractibility factor includes two subtests, Digit Span and Arithmetic. Digit Span appears to measure Gsm. The factorial composition of Arithmetic has been debated (cf. Keith & Witta, 1997; Kranzler, 1997; Phelps et al., 2005). With this in mind, a priori CHC classifications were made for WISC-III subtests (Table 1). Average internal consistency estimates for subtest scores in the norming sample ranged from .69 to .87.

[2] This factorial description for Picture Arrangement is based on findings from the KABC-II Story Completion subtest, a similar subtest, and its association with Gf (Kaufman & Kaufman, 2004; Reynolds et al., 2007).

2.2.3. WISC-IV

The WISC-IV (Wechsler, 2003) includes 15 subtests. Both a four-factor first-order structure (including Verbal Comprehension, Perceptual Reasoning, Processing Speed, and Working Memory) and five-factor first-order structure (with Perceptual Reasoning factor split into Gf and Gv factors) provide good descriptions of the data (Keith, Fine, Reynolds, Taub, & Kranzler, 2006). Eight WISC-IV subtests were used in the current study. A priori CHC subtest classifications are shown in Table 1. Average internal consistency estimates of scores in the norming sample ranged from .81 (for the Comprehension subtest) to .90 (for the Letter–Number Sequencing subtest) for subtests used in this study.

2.2.4. WJ III

The CHC taxonomy was explicitly used throughout the development of the WJ III (Woodcock et al., 2001). The WJ III extended battery measures seven broad CHC abilities, with at least two indicators per factor, but not all of these subtests were administered during the KABC-II validity studies. Nine subtests were used in this study (i.e., subtests of factors represented on the KABC-II): two indicators each of Gc, Gv, Gf, and Gsm, and one indicator of Associative Memory. Those subtests and their predicted CHC classifications are shown in Table 1. Median internal consistency estimates for the scores from these nine subtests ranged from .74 (for the Picture Recognition subtest) to .94 (for the Concept Formation subtest; Schrank, Miller, Wendling, & Woodcock, 2010).

2.2.5. Peabody Individual Achievement Test-Revised/Normative Update (PIAT-R/NU)

The PIAT-R/NU (Markwardt, 1998) is an achievement test, but the General Information subtest assesses general knowledge and was predicted to load on Gc. Internal consistency estimates for scores from this subtest in the norming sample were in the .90s across the age range.

2.3. Study design

One aspect of the validation process for the KABC-II involved studying the correlations of KABC-II test scores with scores from other intelligence batteries (Kaufman & Kaufman, 2004). We used data collected during this process in which samples of children were administered, in counterbalanced order, the KABC-II along with one or more additional intelligence or achievement test batteries. All children were administered the KABC-II. Some children were also administered the WISC-III, others the WISC-III and WJ III, others the WISC-IV, others the PIAT-NU, and others were administered either a mix of these or other tests. Data collection procedures were designed so that scores from the KABC-II and another test battery could be compared. The procedures were not explicitly designed so that data across all batteries would be analyzed simultaneously. Nevertheless, data collection procedures were similar to a planned missingness approach, with the KABC-II subtests serving as the reference indicators because they were administered to everyone across samples. The KABC-II has at least two subtest indicators per CHC common factor, with each of the other batteries used in the current study also containing supposed indicators of some of the five CHC factors measured on the KABC-II (Gc, Gv, Gf, Gsm, and Associative Memory). The two to three subtest indicators per common factor in this study provided a strong reference variable approach. Therefore, data from these different intelligence and achievement test batteries were analyzed simultaneously using CB-CFA with maximum likelihood estimation procedures. We discuss how we handled incomplete data and our plan for analysis in the sections that follow.

2.3.1. Incomplete data

Given the nature of the data collection, every participant in the study had incomplete data. Overall there were 20 different incomplete data patterns. The majority of incomplete data were described by five patterns: (a) KABC-II available and all other variables missing (39%); (b) KABC-II, WISC-III, and WJ III available and other variables missing (22%); (c) KABC-II and PIAT-R/NU available and all others variables missing (17%); (d) KABC-II and WISC-III available and all other variables missing (9%); and (e) KABC-II and WISC-IV available and all other variables missing (14%).

Maximum likelihood estimation procedures were used to deal with incomplete data. Maximum likelihood estimation yields unbiased estimates under the assumption that data are missing completely at random (MCAR) or missing at random (MAR). Said differently, in these situations, nonresponse data may be considered ignorable (Schafer & Graham, 2002). The reference variable approach used here is a type of planned missing design, and "Planned missing values are usually MCAR, but MAR situations sometimes arise" (Schafer & Graham, 2002, p. 152). It may be easiest to conceptualize our data as one large test battery that included all of the variables, with subsets of variables missing because it was never the intention to administer every test battery to every individual. The only measure that was intended to be administered to all of the participants was the KABC-II. Thus, the data in this study may be considered MCAR or MAR (Schafer & Graham, 2002) and the maximum likelihood procedures used in this study are appropriate and should provide unbiased estimates.

Although MAR was considered tenable, additional analyses were conducted to ensure the quality and representativeness of these samples prior to combining raw data into one file. First, participants were generally selected to be representative of the U. S. population, and descriptive analyses showed that the samples were similar in background characteristics (some of these background characteristics, including SES and sex were included in the final analysis). Second, the obtained scores were generally consistent with what would be expected in a large representative sample. Third, tests of measurement invariance for each intelligence battery were performed using the test scores from samples from the KABC-II validity study and from the norming samples for each test. If measurement invariance was mostly supported, it was assumed that the constructs measured in our sample data were measured in the same way that they were in the norming samples. After these considerations were taken into

account, the goal was to combine raw data in one dataset so that they could be analyzed with CB-CFA models. If a subtest was administered, an age-standardized score (raw datum) was entered in the element for that individual. If not, it was coded as missing. All of the scores were then analyzed simultaneously using ML procedures within the Mplus latent variable modeling program (Muthén & Muthén, 1998–2010).

2.3.2. Plan for analysis

The six general steps that were taken during the analysis phase are outlined below, with greater detail provided in the Results section.

1. *Measurement invariance across population and samples.* Tests of factorial invariance were used to investigate measurement invariance with the sample data for tests from the KABC-II validity studies and their parent population data (Meredith, 1993). Invariance provided evidence that the CHC factors were measured in the same way in our samples as they were in the norming sample (i.e., invariance of populations).
2. *WISC-III/IV invariance.* The WISC-IV differed from the WISC-III. Some subtests were dropped and others added, and not all WISC-IV subtests were administered in the KABC-II validity studies. Factorial invariance across the WISC-III/IV batteries was also investigated to assess whether these two instruments could be considered equivalent. Intercept or factor mean differences were expected across samples due to the Flynn effect (Flynn & Weiss, 2007), but these differences could be accommodated in the CB-CFA by using a "test group" dummy variable (discussed in step 5).
3. *Cross-battery CHC measurement model.* A first-order model with the five CHC factors (Gc, Gv, Gf, Gsm, and Associative Memory) was used to model scores from all of the instruments. An acceptable first-order model was developed before introducing a second-order *g* factor.
4. *Second-order cross-battery CHC model.* A second-order *g* factor was introduced to the model.
5. *Cross-battery MIMIC model.* Sex and SES covariates were introduced to control for variance due to between-group mean differences in these variables or constructs. The purpose was not to study the substantive relations, but to obtain the most accurate factor loadings. A failure to control for population heterogeneity in factor models may have led to biased estimates in prior research (Muthén, 1989).
   A second source of potentially unwanted variance was related to expected mean differences on WISC subtests (e.g., the Similarities subtest) across the two versions. Such differences are likely due to the norming date difference (Flynn, 2007). If the covariance structure was invariant, corresponding WISC-III/IV subtest data were merged into one data column. Intercept or mean differences found in the invariance tests across WISC versions in step 2 were accommodated with a dummy variable. Those who were administered a WISC-IV were coded as one, and everyone else was coded as zero. These factor models that include covariates are often referred to as multiple indicator, multiple cause (MIMIC) models (Muthén, 1989).
6. *Monte Carlo study.* A Monte Carlo study was performed to investigate the probability of obtaining biased maximum likelihood parameter estimates and associated standard errors given the number of subtests, incomplete data patterns, and sample size used in this study (Muthén & Muthén, 2002). A population model was used to generate data, a number of samples were generated randomly, and a model was estimated from each of those samples. A first-order, five-factor CHC model similar to the first-order CB-CFA model specified in this study was used as the population model. Factor variances were fixed to one. Population model values are often estimated by using a "best guess" or data from pilot studies. Here, hypothesized values for the population generating model (i.e., factor loadings, residuals, and factor correlations) were obtained from prior within-battery CFA of these measures (Keith & Witta, 1997; Keith et al., 2006; Reynolds et al., 2007) and from CFA models estimated from norming sample data. The pattern and amount of incomplete data were specified to be the same as the five general patterns found in this study. Sample size was fixed to 423. Maximum likelihood estimation was used with 10,000 replications, and these were re-estimated three times with different seed estimates. This analysis provided information regarding power, bias of the parameter estimates and associated standard errors, and the average coverage, which is the proportion of replications for which the 95% confidence interval contained the true parameter. Less than 5% relative parameter bias, less than 10% relative standard error bias, and coverage exceeding .90 were desired (Hoogland & Boomsma, 1998; Muthén & Muthén, 2002).

2.4. Model evaluation

Several model fit indexes were reported as indicators of standalone fit: the root mean square error of approximation with 90% confidence intervals (RMSEA; Steiger & Lind, 1980),[3] the comparative fit index (CFI; Bentler, 1990), and the standardized root mean square residual (SRMR). Values close to or less than .05 are considered good fit for the RMSEA. A CFI value close to or greater than .95 is considered a good fit (Hu & Bentler, 1999). Values <.08 are considered "good," and <.10 acceptable, for the SRMR. The likelihood ratio test ($\Delta\chi^2$) was used for tests of nested models, and $\Delta\chi^2$ and ΔCFI (with ΔCFI > .01 considered meaningful change) were used for tests of factorial invariance (cf. Cheung & Rensvold, 2002; French & Finch, 2006). Last, the sample size-adjusted BIC (aBIC) was reported and may be used to compare models, especially non-nested models. Lower aBIC values indicate a better fitting model.

[3] A correction ($RMSEA \times \sqrt{number\ of\ groups}$) was applied to the RMSEA in multi-group models (Steiger, 1998).

3. Results

3.1. Descriptives for the subtests

Descriptive statistics for the subtests by battery are shown in Table 2. The mean and dispersion of scores were consistent with the general population on most of the subtests. As expected given these well-designed tests and sampling plan, univariate normality was acceptable; all values were well below 2 for skewness and 7 for kurtosis; which are the minimum values often considered problematic for ML estimation (Curran, West, & Finch, 1996).

Table 2
Descriptive statistics for the subtests by battery.

Battery and subtests	*N*	*M*	*SD*	Skewness	Kurtosis
KABC-II					
Atlantis	423	10.20	2.93	−0.22	−0.37
Atlantis Delayed	420	10.06	2.60	−0.19	−0.08
Block Counting	423	9.86	2.88	−0.05	0.19
Expressive Vocabulary	423	9.92	2.91	−0.18	0.70
Gestalt Closure	387	10.02	2.93	0.30	0.09
Hand Movements	423	10.30	2.65	0.11	−0.01
Number Recall	423	10.65	2.91	−0.06	−0.18
Pattern Reasoning	423	10.21	2.82	0.16	−0.22
Rebus	423	10.32	3.06	−0.02	−0.21
Rebus Delayed	417	10.10	2.91	−0.27	−0.41
Riddles	423	10.43	2.84	−0.04	−0.13
Rover	423	10.38	3.08	0.03	0.03
Story Completion	423	9.97	2.85	−0.24	0.31
Triangles	423	10.57	2.67	−0.35	−0.38
Verbal Knowledge	423	10.09	2.81	−0.06	−0.03
Word Order	423	9.94	2.61	0.07	0.19
WISC-III					
Arithmetic	127	10.77	2.97	0.14	−0.23
Block Design	127	11.12	2.89	−0.03	1.35
Comprehension	127	10.36	3.77	−0.23	−0.06
Digit Span	114	9.96	2.99	0.27	−0.76
Information	127	10.78	3.41	−0.38	−0.30
Object Assembly	127	10.29	2.93	−0.19	0.60
Picture Arrangement	127	10.69	3.43	0.19	−0.18
Picture Completion	127	10.61	3.10	−0.35	1.29
Similarities	127	11.22	3.18	−0.77	0.78
Vocabulary	127	10.46	3.51	−0.37	0.47
WISC-IV					
Block Design	58	9.48	3.24	−0.06	0.40
Comprehension	58	9.34	2.92	−0.46	−0.34
Digit Span	58	9.45	2.81	−0.09	−0.01
Letter–Number Sequencing	58	8.88	3.14	−0.94	0.85
Matrix Reasoning	58	9.78	3.10	0.20	−0.92
Picture Concepts	58	9.79	3.22	0.42	0.08
Similarities	58	8.91	3.04	−0.14	−0.68
Vocabulary	58	9.29	3.45	−0.22	−0.74
WJ III					
Verbal Comprehension	91	102.11	14.60	−0.69	0.53
Visual–Auditory Learning	91	94.23	19.89	−0.43	1.91
Spatial Relations	91	100.47	11.32	−0.68	1.45
Concept Formation	91	105.18	13.93	−0.06	0.62
Numbers Reversed	91	100.43	14.33	−0.02	0.61
Auditory Working Memory	90	105.19	13.91	0.38	−0.02
General Information	91	97.87	16.49	−0.30	0.28
Picture Recognition	91	100.64	12.55	0.07	3.13
Analysis-Synthesis	89	102.85	17.11	−0.29	0.56
PIAT-R/NU					
General Information	73	102.12	14.80	−0.30	0.41

Note. WISC-III = Wechsler Intelligence Scale for Children—Third Edition (Wechsler, 1991); WISC-IV = Wechsler Intelligence Scale for Children—Fourth Edition (Wechsler, 2003); KABC-II = Kaufman Assessment Battery for Children—Second Edition (Kaufman & Kaufman, 2004); WJ III = Woodcock-Johnson III Tests of Cognitive Abilities (Woodcock et al., 2001); PIAT-R/NU = Peabody Individual Achievement Test-Revised/Normative Update (Markwardt, 1998).

3.2. Measurement invariance across populations and samples

Meredith (1993) outlined steps for using factorial invariance to test for measurement invariance via MG-CFA. Configural invariance is tested first by associating the same set of subtests with the same first-order factor across groups. Next, equality constraints on corresponding parameter estimates across groups are applied, and the fit of each these more constrained models is compared to the fit of a less constrained model. Weak factorial invariance (equal factor loadings), strong factorial invariance (equal factor loadings and intercepts), and strict factorial invariance (equal factor loadings, intercepts, and residual variances) were tested. Factor mean differences across groups were estimated while testing for strong and strict factorial invariance.[4] The most stringent test, a test of strict factorial invariance, was applied after an acceptable configural invariance model was developed. If strict factorial invariance was untenable, follow up tests were performed to locate the source of strain on model fit and to find a model exhibiting weak or strong factorial invariance. Results for all of the invariance tests, by test battery, are shown in Table 3.

First, invariance tests were performed to determine if the KABC-II factors were measured in the same way in the sample data from the validity studies ($n = 423$) as they were in the norming sample. A random sample of 500 participants, 6 to 16 years in age, was selected from the KABC-II norming sample data so that the sample size was similar across groups. The Configural Invariance model for the KABC-II was a first-order, five-factor measurement model consistent with the test structure. Configural invariance was tenable (see Table 3, Model 1a). The fit of the Strict Factorial Invariance (see Table 3, Model 1b) model was statistically significantly worse ($p = .007$) than the Configural Invariance according to the likelihood ratio test. Given the number of constraints that were applied and lack of change in other indexes (ΔCFI $= .004$), many researchers may consider this result as evidence for invariance. Nevertheless, follow-up tests were performed. Corresponding factor loadings (see Table 3, Model 1c) and residuals (see Table 3, Model 1d) were invariant using strict criteria. The CHC factors were measured in the same way in our sample as they were in the KABC-II norming sample.

Second, factorial invariance of the WISC-III was investigated. A covariance matrix that was averaged across age in previous research was used as input for the WISC-III norming sample model (Keith & Witta, 1997). A four-factor model (including Verbal Comprehension, Perceptual Organization, Processing Speed, and Short-Term Memory) was specified. Strict factorial invariance was tenable (see Table 3, Model 2b).

Third, factorial invariance of the WISC-IV was tested. An averaged WISC-IV norming sample covariance matrix created by Keith et al. (2006) was used as WISC-IV population data. The structure used for the WISC-IV invariance models was consistent with the structure found in the research by Keith et al. (2006). Block Design was the only indicator of the Gv factor. The residual variance for this subtest was fixed using the average reliability estimate for the subtest scores provided in the manual (Keith, 2006). The numbers of cases was fixed to 100 to be more consistent with sample size in the KABC-II validity study. Strict factorial invariance was tenable (see Table 3, Model 3b).

Last, invariance of the WJ III was tested across the KABC-II validity sample and a random sample of 99 participants, 6 to 16 years in age, drawn from the WJ III norming sample. The measurement model for the WJ III dataset included four first-order factors (Gc, Gv, Gf, and Gsm), each indicated by two subtests. Strict factorial invariance was untenable (see Table 3, Model 4b). Lack of invariance was due primarily to the Analysis–Synthesis factor loading and the Spatial Relations and Picture Recognition residual variances. A Partial Strict Factorial Invariance model that allowed for these to be estimated freely within groups was tenable (see Table 3, Model 4c).

Partial strict factorial invariance was not a major concern, but it may have implications for the generalization of the findings. Because of this finding, all CB-CFA models were run with and without the Gf and Gv subtests from the WJ III. No substantively important differences in parameter estimates were noted, including second-order factor loadings. Therefore, results with all of the available WJ III subtests are reported, while recognizing that only partial invariance of the loadings and residuals was supported. If interested in the results for the CB-CFA models estimated without the WJ III subtests, please contact the first author. It should be noted that partial invariance related to factor loadings has been shown to have negligible influence on structural parameters in other research (Schmitt, Golubovich, & Leong, 2011).

In general, tests of factorial invariance indicated that invariance across populations and samples was tenable. The common factors were measured in the same way in these smaller samples as they were in the norming samples. Other than one, all corresponding factor loadings were the same, indicating that the constructs were similar across samples. These results suggest that subsequent findings can be generalized to the larger norming samples.

3.3. WISC-III/IV invariance

A MG-CFA model was estimated with each WISC battery representing a group. If a WISC-III subtest was not included in either the WISC-IV revision or not administered as part of the KABC-II validity study, the subtest was represented by a latent variable (i.e., ovals were substituted for rectangles for the subtests) in the WISC-IV model with zero variance (i.e., missing data) as shown in Fig. 1 (McArdle, 1994; Wothke, 2000). Likewise, new subtests included in the WISC-IV revision were represented in the same way in the WISC-III group. Four first-order factors were specified. Measurement invariance was investigated by constraining corresponding factor loadings, intercepts (while freely estimating latent factor means in the non-reference group),

[4] A reference group was specified, with each latent factor mean fixed to zero in that group. Latent factor means in the other group were estimated freely; thus, those values represented the corresponding latent mean difference from the reference group.

Table 3
Tests of factorial invariance across sample and standardization data for the test batteries and across WISC-III and WISC-IV versions.

	$\chi 2$ (*df*)	*p*	$\Delta\chi 2$ (Δdf)	*p*	CFI	Adj. RMSEA
KABC-II validity sample and test norming sample						
KABC-II						
1a. Configural Invariance	251.43(180)	<.01			.99	.03
1b. Strict Factorial Invariance	316.45(220)	<.01	65.02(40)	.01	.98	.03
1c. Weak Factorial Invariance (Factor Loadings)[a]	269.31(193)	<.01	17.88(13)	.16	.99	.03
1d. Factor Loadings + Residuals equal	294.10(209)	<.01	24.79(16)	.07	.99	.03
WISC-III						
2a. Configural Invariance	92.14(96)	.59			1.00	<.01
2b. Strict Factorial Invariance	121.75(124)	.54	29.62(28)	.38	1.00	<.01
WISC-IV						
3a. Configural	58.60(98)	.99			1.00	<.01
3b. Strict Factorial Invariance	63.97(106)	.99	5.37(8)	.72	1.00	<.01
WJ III						
4a. Configural Invariance	43.44(28)	.03			.97	.08
4b. Strict Factorial Invariance	104.44(44)	<.01	61.00(16)	<.01	.88	.13
4c. Partial Strict Factorial Invariance[b] (Analysis–Synthesis factor loading and the Spatial Relations and Picture Recognition residuals freed)	71.08(41)	.02	27.69(13)	.01	.94	.08
WISC-III & IV						
5a. Configural Invariance	100.35(76)				.97	.06
5b. Strict Factorial Invariance	157.90(90)	<.01	57.55(14)	<.01	.92	.08
5c. Strong Factorial Invariance [c]	122.82(82)	<.01	22.47(6)	<.01	.92	.08
5d. Partial Strong Factorial Invariance[c] (Similarities intercept freed)	106.70(81)	.03	6.35(5)	.27	.97	.06

Note. Compare all models with previously listed model unless noted otherwise. WISC-III = Wechsler Intelligence Scale for Children—Third Edition (Wechsler, 1991); WISC-IV = Wechsler Intelligence Scale for Children—Fourth Edition (Wechsler, 2003); KABC-II = Kaufman Assessment Battery for Children—Second Edition (Kaufman & Kaufman, 2004); WJ III = Woodcock-Johnson III Tests of Cognitive Abilities (Woodcock et al., 2001); PIAT-R/NU = Peabody Individual Achievement Test-Revised/Normative Update (Markwardt, 1998).

[a] Compare with Model 1a.
[b] Compare with Model 4a.
[c] Compare with model 5a.

and residual variances equal. The fit of the Configural Invariance model was acceptable (see Table 3, Model 5a). There was degradation in fit in the Strict Factorial Invariance model (see Table 3, Model 5b). Differences in the Similarities subtest intercept were expected (Flynn & Weiss, 2007). Thus, this intercept was freed, and indeed partial strong factorial invariance was tenable (see Table 3, Models 5c and 5d). Mean differences in the Similarities subtest scores were not accounted for by factor mean differences.

Given invariance of the covariance structure, scores from five subtests that were administered in both the WISC-III and WISC-IV (i.e., Block Design, Digit Span, Similarities, Comprehension, and Vocabulary) were merged. For example, the Similarities subtest scores from the WISC-III and WISC-IV were included in the same data column. A dummy variable with one group representing the WISC-IV group was eventually included in the CB-CFA higher order models, which accommodated the intercept difference in Similarities (Meredith & Teresi, 2006). WISC-IV subtest data that were not merged were from the Picture Concepts, Matrix Reasoning, and Letter–Number Sequencing subtests. WISC-III subtest data not merged were from the Picture Arrangement, Object Assembly, Picture Completion, Arithmetic, and Information subtests. These subtests were either not included in the WISC-IV (e.g., Object Assembly) or their corresponding WISC-IV subtests were not administered in the KABC-II validity study (e.g., Arithmetic). WISC-III and IV subtests are referred to as "WISC" subtests from this point on.

3.4. Cross-battery CHC measurement model

A first-order CB-CFA model with five factors was specified to account for the covariances among all of the subtests across batteries. A priori configuration of these subtests by factor is shown in Table 1: Gc was indicated by 12 subtests, Gv by 9 subtests, Gf by 9 subtests, Associative Memory by 5 subtests, and Gsm by 8 subtests. A priori assignments to factors were based on CHC-referenced studies of individual batteries along with descriptions of CHC narrow and broad abilities (e.g., Flanagan et al., 2012).

Subtest residual variances from the KABC-II Delayed Recall subtests were freed to correlate with residuals from the initial measurement (e.g., Atlantis Delayed with Atlantis). The Gsm factor included both working memory and memory span tasks; residuals were correlated across the three memory span tasks Word Order, Digit Span, and Number Recall, essentially creating a Memory Span factor. As a result, the Gsm factor thus represented working memory more than memory span.[5] Four subtests,

[5] Alternatively, it would have been possible to model these as a narrow ability factor subsumed by a broad ability, but that was beyond the focus of this study.

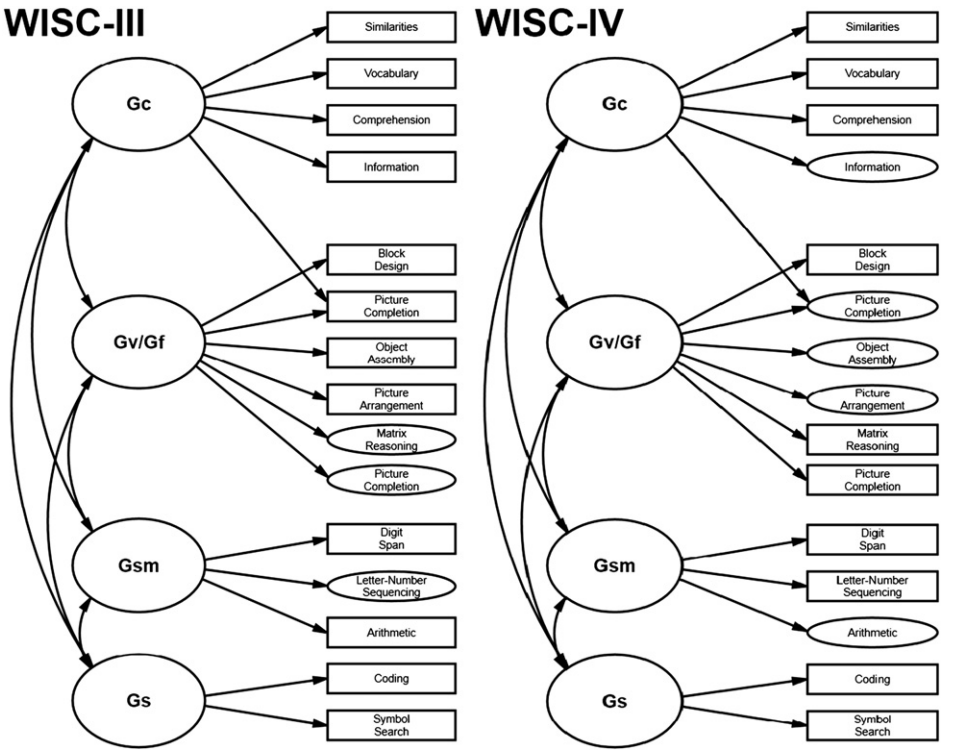

Fig. 1. WISC-III and WISC-IV multi-group conceptual model. *Note*. Corresponding factor loadings and intercepts were constrained equal across WISC groups in the Strong Factorial Invariance model. The WISC-III factor means were fixed to zero, and the WISC-IV factor means were estimated. WISC-III = Wechsler Intelligence Scale for Children—Third Edition (Wechsler, 1991); WISC-IV = Wechsler Intelligence Scale for Children—Fourth Edition (Wechsler, 2003).

Table 4
Model fit statistics for first-order, second-order, and second-order with covariates cross-battery models.

Model	$\chi 2$ (df)[a]	$\Delta\chi 2$ (Δdf)	p	CFI	RMSEA (90% CI)	SRMR	aBIC
CHC measurement models							
1. Initial model	1033.90(641)			.92	.038 (.034–.042)	0.09	44944.8
2. Cross-loadings deleted	1036.97(644)	3.07(3)	.38	.92	.038 (.034–.042)	0.09	44939.3
3. Correlated specifics	1002.97(641)	34.00(3)	<.01	.93	.037 (.032–.041)	0.09	44913.9
4. Pic Rec on MA + residuals	988.69(640)	14.28(1)	<.01	.93	.036 (.031–.040)	0.09	44902.5
5. Pic Rec on MA only	989.34(641)	0.65(1)	.42	.93	.036 (.031–.040)	0.09	44900.3
Second-order models							
6. Second-order model	1001.10(646)	11.76(5)	.04	.93	.036 (.032–.040)	0.09	44897.7
7. Second-order Gf = 0	1002.55(647)	1.45(1)	.23	.93	.036 (.032–.040)	0.09	44896.3
Second-order with covariates							
8. Gf unique = 0[b]	1125.64(752)			.93	.034 (.030–.038)	0.09	46843.2
9. Gf unique estimated	1125.13(751)	.51 (1)	.48	.93	.034 (.030–.038)	0.09	46845.6

Note. Compare with previously listed model unless otherwise noted.
Pic Rec = WJ III Picture Recognition; Gf = Fluid/Novel Reasoning; MA = Associative Memory.
[a] All = $p < .05$ in this column.
[b] Do not compare with previous models.

KABC-II Gestalt Closure, WISC Picture Completion, KABC-II Hand Movements, and WISC Arithmetic, were cross-loaded (across two factors) because of prior research findings (e.g., Keith et al., 2006; Reynolds et al., 2007).

Results of the CB-CFA of the measurement model are presented in the top portion of Table 4. The fit of the Initial CB-CFA model was acceptable (see Table 4, Model 1). The value for the RMSEA was considered excellent, but the CFI was less than desired, and the SRMR, although acceptable, was not considered "good." Three factor loadings were not statistically significant or substantial: WISC Picture Completion on Gv, KABC-II Hand Movements on Gf, and WISC Arithmetic on Gf. These three were cross-loadings. The KABC-II Gestalt Closure factor loadings were small in magnitude yet statistically significant. All of the remaining standardized loadings were substantial in magnitude (> .40) and statistically significant. The three statistically nonsignificant factor loadings were deleted (see Table 4, Model 2).

Although global fit was acceptable, modification indexes were examined for potential sources of local misfit. Residual (specific-factor) correlations were expected because of the number of indicators, with some subtests measuring common CHC narrow abilities. The investigation of every possible correlated specific was not the purpose of this study, but a few modification indexes were examined to investigate the influence of these specific correlations on model parameters (Heene, Hilbert, Freudenthaler, & Bühner, 2012). Three of the largest modification indexes were related to correlated specifics involving the KABC-II Story Completion and KABC-II Gestalt Closure subtests, the WJ III Concept Formation and KABC-II Pattern Reasoning subtests, and the WJ III Picture Recognition and Spatial Relations subtests. Two correlated specifics appeared substantively reasonable, as the ability to see the larger picture or whole is important for success on the KABC-II Story Completion and KABC-II Gestalt Closure subtests and deductive reasoning is required for both the WJ III Concept Formation and the KABC-II Pattern Reasoning subtests. The third correlated specific was less clear. Gv is measured by both the WJ III Picture Recognition and WJ III Spatial Relations subtests, so perhaps this was a method effect or a sample specific finding as the residual variances of these subtests were not invariant in previous analysis. Another model was estimated with these three additional correlated specifics, resulting in a statistically significant improvement in model fit (see Table 4, Model 3). The correlated specifics did not influence other model parameters. No remaining modification indexes for specific correlations were greater than 10 ($p < .001$). Additional modifications related to correlated specifics may have improved fit, but testing all of these was beyond the purpose of this study.

One large modification index (13.64) indicated that the WJ III Picture Recognition subtest should load on the Associative Memory factor, which was also evident in the previous model. Picture Recognition was found to load on a Glr factor in a recent study (Kaufman et al., 2012). Allowing for this cross-loading improved model fit, and the loading of Picture Recognition on Gv was no longer statistically significant (see Table 4, Model 4). The cross-loading was deleted (see Table 4, Model 5). There were no additional large modification indexes.

With relatively few changes to the a priori model specification, the global fit indexes and local modification indexes were indicative of an acceptable to well-fitting measurement model. The general fit of this model suggests that factor indicators when analyzed in combination indeed loaded on their predicted CHC factors when those factor assignments were based on previous single-battery analyses and the CHC taxonomy. Thus, the CHC-based factors were invariant across intelligence batteries; in the words of Tucker (1958), the factors indeed "transcended" their individual batteries. The factorial composition of only one subtest, WJ III Picture Recognition, was not well described by the predicted CHC factor. The fact that CHC model based on CHC predictions fit reasonably well also provides support for CHC theory as a way of understanding the structure of cognitive tests. Model 5 from Table 4, with standardized estimates, is shown in Fig. 2.

3.5. Second-order cross-battery CHC model

The fit of the second-order CB-CFA model was also acceptable. Fit was statistically significantly worse compared to final first-order CHC model according to the likelihood ratio test but not the aBIC (see Table 4, Model 6). Second-order factor loadings were large: Gsm = .79, Associative Memory = .80, Gv = .80, and Gc = .83. The loading of Gf on *g* was .96, and the Gf first-order unique variance was not statistically significantly different from zero, suggesting that the two factors were perfectly correlated. Indeed, there was not a statistically significant degradation in fit when the Gf unique variance was fixed to zero (see Table 4, Model 7). Gf and *g* were statistically indistinguishable.

3.6. Cross-battery MIMIC model

Sex and SES covariates were used to investigate the effects of population heterogeneity on the factor loadings. In addition a dummy-coded "test battery" variable was included to account for an intercept difference (i.e., extension of range) between the Similarities subtest across WISC-IV and WISC-III versions. Similarities was regressed on this test-battery covariate, and *g* and the first-order factors regressed on all three covariates. Because a model in which all of the first-order factors and *g* were regressed simultaneously on the covariates would be under-identified, one broad ability was not allowed to be regressed on the covariates, and different combinations were estimated (e.g., Gsm was not regressed on the covariates in one model and Associative Memory was not regressed in another). In all of these models, the first-order Gf unique variance was not statistically significantly different from zero, and the loading approached one (.99) when Gf was regressed on SES. Therefore, the Gf unique variance was fixed to zero, which effectively collapsed this factor with *g*. With the Gf unique variance fixed to zero, all of the common factors could be regressed on the covariates simultaneously and the model was identified. This model with the Gf unique variance fixed to zero and all other common factors regressed on the covariates had acceptable fit (see Table 4, Model 8). All parameter estimates in this model were reasonable. A final model was estimated in which the Gf unique variance was not fixed to zero (and Gf was not regressed on the covariates). Fit did not improve, again indicating that Gf and *g* were statistically inseparable ($r = .98$; see Table 4, Model 9).

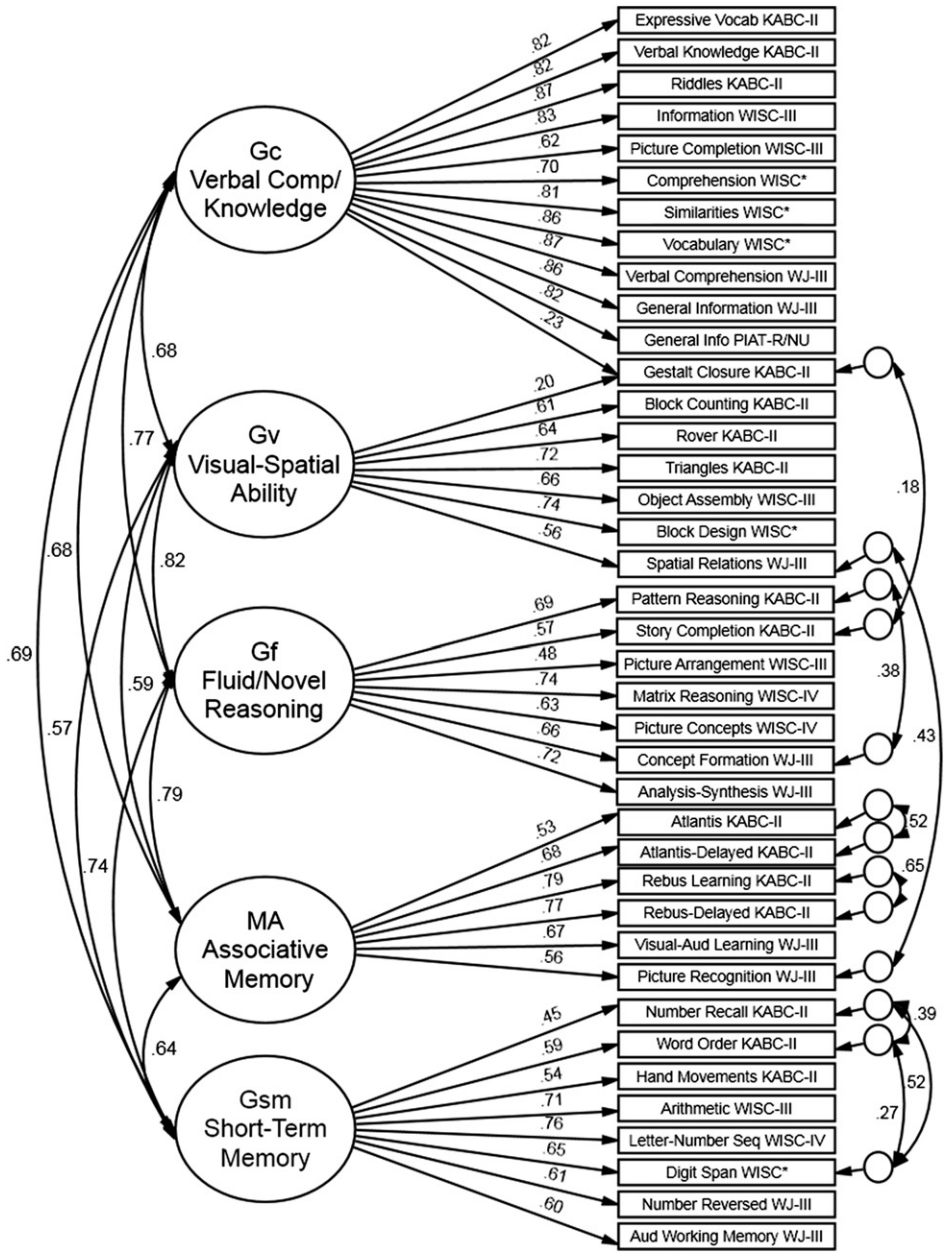

Fig. 2. First-order CHC measurement model with factor correlations and standardized factor loadings. *Note*. Measurement residual variances were deleted from the figure unless they were correlated. WISC-III = Wechsler Intelligence Scale for Children—Third Edition (Wechsler, 1991); WISC-IV = Wechsler Intelligence Scale for Children—Fourth Edition (Wechsler, 2003); KABC-II = Kaufman Assessment Battery for Children—Second Edition (Kaufman & Kaufman, 2004); WJ III = Woodcock-Johnson III Tests of Cognitive Abilities (Woodcock et al., 2001); PIAT-R/NU = Peabody Individual Achievement Test-Revised/Normative Update (Markwardt, 1998). WISC* indicates tests included on both versions of the WISC and administered in the KABC-II validity study. Untimed scores were used for the KABC-II Pattern Reasoning, Story Completion, and Triangles subtests.

Model 9 from Table 4 was used to report the parameter estimates shown in Fig. 3.[6] All first-order factor loadings were statistically significant at the $p < .05$ level, and all of the standardized first-order factor loadings were greater than .40 except for the KABC-II Gestalt Closure subtest. These loadings were similar to the first-order standardized loadings shown in Fig. 2.

Second-order factor loadings were generally large: Gsm = .75, Gf = .98, Associative Memory = .82, Gv = .76, and Gc = .75. There were some differences in the second-order factor loadings in the model with covariates (conditional model), when compared to the second-order model without covariates (unconditional model). Associative Memory (.01) and Gf (.02) loadings

[6] Although the fit did not improve in Model 9 these estimates were reported because they may be compared to the second-order model without covariates. The Gf loading may also be interpreted as not being statistically significantly different from one in any of the higher-order models.

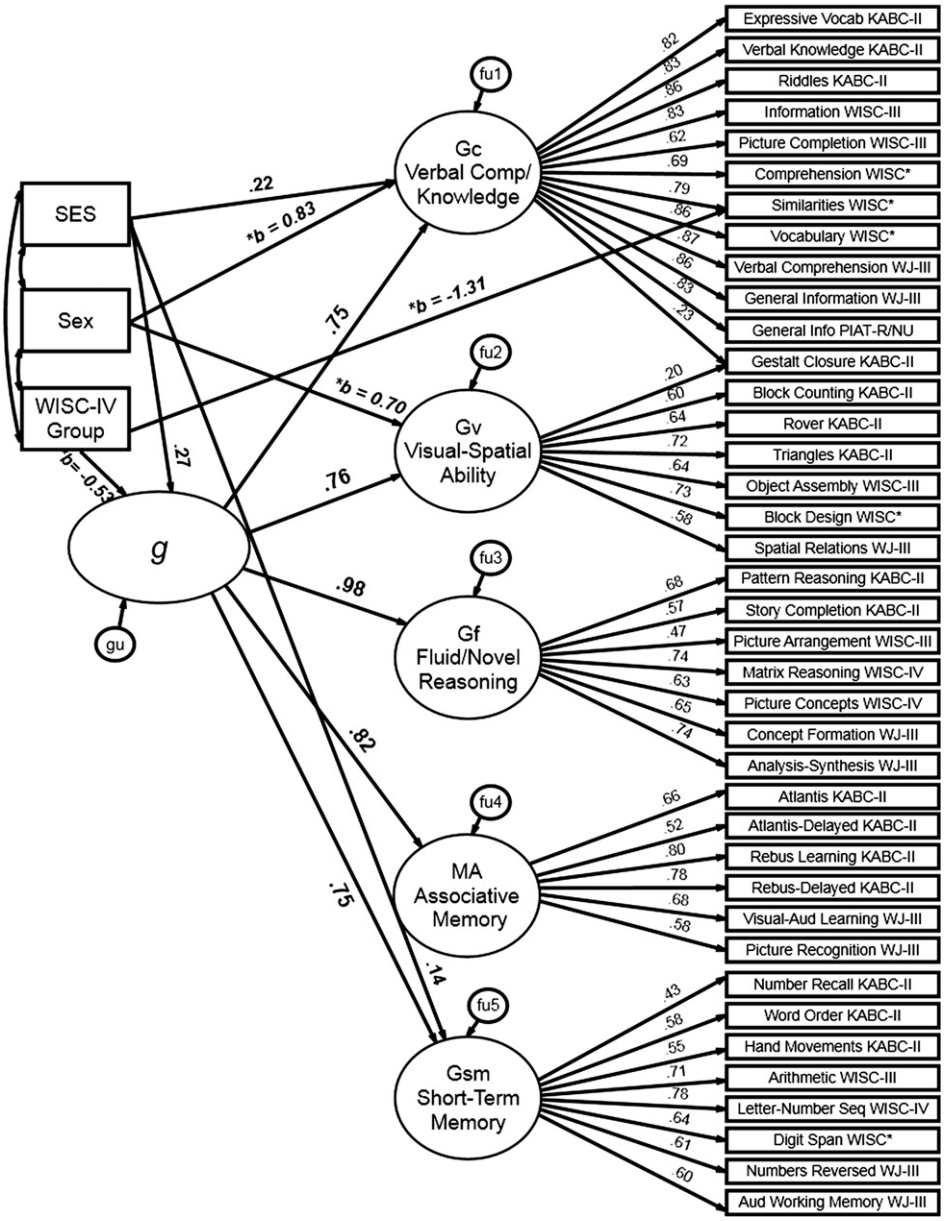

Fig. 3. Second-order CHC model with covariates. *Note*. Measurement residual variances and correlations were deleted from the figure. All loadings and effects are standardized, unless indicated by an *b. WISC-III = Wechsler Intelligence Scale for Children—Third Edition (Wechsler, 1991); WISC-IV = Wechsler Intelligence Scale for Children—Fourth Edition (Wechsler, 2003); KABC-II = Kaufman Assessment Battery for Children—Second Edition (Kaufman & Kaufman, 2004); WJ III = Woodcock-Johnson III Tests of Cognitive Abilities (Woodcock et al., 2001); PIAT-R/NU = Peabody Individual Achievement Test-Revised/Normative Update (Markwardt, 1998). The WISC-IV dummy variable indicates those individuals who took the WISC-IV. WISC* indicates tests included on both versions of the WISC and administered in the KABC-II validity study. Positive values for the sex dummy variable indicate boys were higher. Untimed scores were used for the KABC-II Pattern Reasoning, Story Completion, and Triangles subtests.

increased slightly, but Gsm (−.03), Gc (−.08), and Gv (−.05) decreased. On average, there was a .04 change in the absolute values of the second-order *g* loadings.

Last, the parameters from this higher-order (MIMIC) model in Fig. 3 were transformed using residualized common factor variances to estimate the relative effects of the common factors on the subtests (Schmid & Leiman, 1957). Estimates for the higher-order models without and with the covariates are shown on the left and right sides of Table 5, respectively. Both *g* and broad CHC common factors, except Gf, had substantial effects on subtest scores. Both standardized effects and the proportion of subtest variance explained (in parentheses) by the common factors are reported in Table 5. Although researchers often interpret the proportion of variance explained in the subtests, standardized effects are proportionally similar to each other and more appropriate for relative interpretations (Darlington, 1990). As shown on the right side of Table 5, the effect of *g* (represented as a *g*

Table 5
Relative influence of g and CHC factors on subtests in the no covariate and covariate models.

Test battery	CHC factor	Subtest	No covariate model[a]		Covariate model[b]	
			g effects (% variance explained)	CHC factor effects (% variance explained)	*g* effects (% variance explained)	CHC factor effects (% variance explained)
KABC-II	Gc	Expressive Vocabulary	.68(47)	.45(21)	.59(34)	.43(18)
KABC-II	Gc	Verbal Knowledge	.68(47)	.45(21)	.59(35)	.43(19)
KABC-II	Gc	Riddles	.72(51)	.48(23)	.61(38)	.45(20)
WISC-III	Gc	Information	.69(47)	.46(21)	.59(35)	.43(18)
WISC-III	Gc	Picture Completion	.52(27)	.34(12)	.44(19)	.32(10)
WISC[c]	Gc	Comprehension	.58(34)	.39(15)	.49(24)	.36(13)
WISC[c]	Gc	Similarities	.67(45)	.45(20)	.56(31)	.41(17)
WISC[c]	Gc	Vocabulary	.72(52)	.48(23)	.61(38)	.45(20)
WJ III	Gc	Verbal Comprehension	.72(51)	.48(23)	.62(38)	.45(20)
WJ III	Gc	General Information	.72(51)	.48(23)	.61(37)	.44(20)
PIAT-R/NU	Gc	General Information	.69(47)	.46(21)	.59(35)	.43(19)
KABC-II	Gc	Gestalt Closure	.20(04)	.13(02)	.16(03)	.12(01)
***M* =**			**.63(42)**	**.42(19)**	**54(31)**	**39(16)**
KABC-II	Gv	Gestalt Closure	.15(02)	.11(01)	.14(02)	.12(01)
KABC-II	Gv	Block Counting	.49(24)	.36(13)	.43(19)	.35(12)
KABC-II	Gv	Rover	.51(26)	.38(15)	.46(21)	.38(14)
KABC-II	Gv	Triangles	.57(33)	.43(18)	.52(27)	.42(18)
WISC	Gv	Object Assembly	.52(27)	.39(15)	.46(21)	.38(14)
WISC[c]	Gv	Block Design	.59(35)	.44(19)	.53(28)	.43(18)
WJ III	Gv	Spatial Relations	.46(21)	.34(12)	.42(17)	.34(11)
***M* =**			**.47(24)**	**.35(13)**	**.42(19)**	**.34(13)**
KABC-II	Gf	Pattern Reasoning	.66(44)	.18(03)	.63(40)	.14(02)
KABC-II	Gf	Story Completion	.54(29)	.15(02)	.53(28)	.12(01)
WISC-III	Gf	Picture Arrangement	.47(22)	.13(02)	.43(19)	.10(01)
WISC-IV	Gf	Matrix Reasoning	.74(51)	.20(04)	.69(47)	.15(02)
WISC-IV	Gf	Picture Concepts	.63(37)	.17(03)	.59(34)	.13(02)
WJ III	Gf	Concept Formation	.63(40)	.17(03)	.61(37)	.14(02)
WJ III	Gf	Analysis–Synthesis	.70(48)	.19(04)	.69(47)	.15(02)
***M* =**			**.62(39)**	**.17(03)**	**.60(36)**	**.13(02)**
KABC-II	MA	Atlantis	.42(17)	.32(10)	.40(16)	.30(09)
KABC-II	MA	Atlantis Delayed	.54(29)	.41(16)	.52(27)	.39(15)
KABC-II	MA	Rebus	.62(38)	.47(22)	.61(37)	.46(21)
KABC-II	MA	Rebus Delayed	.63(39)	.48(23)	.62(39)	.47(22)
WJ III	MA	Visual–Auditory Learning	.54(29)	.41(17)	.52(27)	.40(16)
WJ III	MA	Picture Recognition	.45(20)	.34(10)	.45(20)	.34(11)
***M* =**			**.53(29)**	**.40(17)**	**.52(28)**	**.39(16)**
KABC-II	Gsm	Number Recall	.35(12)	.27(07)	.31(09)	.25(06)
KABC-II	Gsm	Word Order	.46(21)	.36(13)	.41(17)	.34(12)
KABC-II	Gsm	Hand Movements	.43(18)	.34(11)	.39(15)	.33(11)
WISC-III	Gsm	Arithmetic	.56(31)	.44(19)	.51(31)	.42(18)
WISC-IV	Gsm	Letter–Number Sequencing	.61(37)	.48(23)	.63(40)	.46(22)
WISC[c]	Gsm	Digit Span	.51(26)	.40(16)	.46(21)	.38(14)
WJ III	Gsm	Numbers Reversed	.48(23)	.37(14)	.44(19)	.36(13)
WJ III	Gsm	Auditory Working Memory	.46(22)	.36(13)	.43(19)	.36(13)
***M* =**			**.48(24)**	**.38(15)**	**.44(20)**	**.36(14)**

Note. Untimed scores were used for the KABC-II Pattern Reasoning, Story Completion, and Triangles subtests. Gc = Verbal Comprehension-Knowledge; Gv = Visual–Spatial Ability; Gf = Fluid/Novel Reasoning; MA = Associative Memory; Gsm = Short-Term Memory.
[a] Estimates from Table 4, Model 6.
[b] Estimates from Table 4, Model 9.
[c] WISC-III/IV combined.

loading) on the KABC-II Block Counting subtest scores is .43, whereas the effect of Gv on Block Counting scores is .35. A one standard deviation increase in *g* is related to a .43 standard deviation increase in Block Counting scores, and a one standard deviation increase in Gv, independent of *g*, is related to a .35 standard deviation increase in Block Counting scores. When interpreted in this light, it is clear that both *g* and Gv are important for success on the Block Counting task. Last, as shown in the covariate model estimates on the right side of Table 5, the covariates altered some of the common factor effects on the subtest scores. In particular, *g* effects on Gc subtests were reduced. Gf subtests, on average, had the strongest *g* loadings when heterogeneity was controlled. The two subtests with the strongest *g* loadings were the Analysis–Synthesis subtest from the WJ III and Matrix Reasoning from the WISC-IV. When heterogeneity was not controlled, Gc subtests, on average, had stronger *g* loadings.

Last, some of the covariates had statistically significant effects. The WISC-IV dummy variable showed a negative and statistically significant effect on *g* ($b = -0.53$, $p = .008$ $\beta = -.14$), meaning that the WISC-IV group's *g* was lower compared to the WISC-III

group, which is consistent with a Flynn effect.[7] In addition, the WISC-IV group showed a lower score on the Similarities subtest ($b = -1.31$, $p < .001$, $\beta = -.15$) even with g differences controlled. Sex had a statistically significant effect on Gc and Gv. Controlling for differences in g, boys scored higher on Gc ($b = 0.83, p < .001, \beta = .18$) and Gv ($b = 0.70, p < .001, \beta = .20$) than did girls. Last, there was a positive and statistically significant effect of SES on g ($b = 0.36, p < .001, \beta = .27$), and Gc ($b = 0.55, p < .001$, $\beta = .22$) and Gsm ($b = .18, p = .027, \beta = .14$) beyond g. Together the covariates explained 10% of the g variance.

3.7. Monte Carlo study

The results of the Monte Carlo study supported the use of our methods given the sample size and patterns of and amount of incomplete data (Muthén & Muthén, 2002). The magnitudes for relative parameter bias were all <5%, with the vast majority <1% (Hoogland & Boomsma, 1998). Based on the pattern and amount of incomplete data, the parameter estimates were considered of high quality. The same could be said for the standard errors. The magnitudes of the relative standard error bias were all <5%, suggesting precision in the obtained estimates. Coverage for all of the factor loadings and covariances was acceptable, with values ranging from .94 to .95, all very close to the correct value of .95. Coverage for the residual variances ranged from .91 to .95, which was also acceptable. Last, power estimates for all of the parameters except for one were excellent; power was one for all but five. The only estimate that had power less than .80 was the WISC Picture Completion cross-loading on Gc (power = .67).

4. Discussion

A reference variable methodology was used to perform cross-battery confirmatory factor analysis (CB-CFA) on subtests from five different intelligence and achievement batteries to address two primary aims. The first aim was to examine the utility of the CHC taxonomy by investigating the invariance of CHC broad ability factors across different test batteries. CHC theory and prior research were used to predict which subtest loaded on the first-order CHC factor. All but one of the 39 subtests, from five different test batteries that were developed by different test authors, loaded on their predicted CHC factors. Findings thus support CHC theory and its use as a taxonomy for test development.

The second aim of the study was to investigate second-order factor loadings. Of specific interest was the magnitude of the loadings, whether Gf and g were distinguishable, and the influence of population heterogeneity on the loadings. The loadings were similar to those of a previous study in which CHC theory was used to automatically generate items (Arendasy et al., 2008), Gf and g were statistically indistinguishable (Gustafsson, 1984; Kvist & Gustafsson, 2008), and population heterogeneity related to sex and SES influenced the magnitude of second-order factor loadings. In addition, the loadings of Gc subtests on g, in particular, were reduced when the covariates were included. Gf subtests, on average, had the strongest g loadings.

Last, a reference variable approach, which may be applied in planned missingness research, was an efficient way of analyzing data from different intelligence batteries (McArdle, 1994). Breadth of measurement was enhanced substantially. The findings and implications with regard to the two primary aims are discussed in more detail in the sections that follow.

4.1. Cattell–Horn–Carroll taxonomy

"Do factors transcend a particular battery" (Tucker, 1958, pg. 112)? This research shows that many CHC factors do. Testing a priori predictions is essential to science. Here, predictions based on CHC theory and prior studies were used to classify subtests from five different test batteries onto CHC factors. The predictions were straightforward and accurate, even when the subtests were from an intelligence battery that was not specifically based on the CHC taxonomy (e.g., WISC-III and WISC-IV). These findings are of practical and theoretical import, although they are interrelated. From a practical standpoint, the generally well-defined CHC taxonomy should continue to serve as a useful guide that can be used for intelligence test development, assessment, and interpretation (Schneider & McGrew, 2012). From a theoretical standpoint, these findings build upon the work of Carroll's (1993) synthesis because scores from several different intelligence batteries were factor analyzed jointly, providing strong evidence that the CHC factors that arise within intelligence batteries also arise across intelligence batteries. Such invariance of these factors is important, as early factor analysts (Thurstone, 1947; Tucker, 1958) were concerned about the stability of factors. If the factorial composition of subtests changes each time they are moved from one battery to another, the utility of those factors and the tests that purport to measure them would be questioned. Here it was shown that the factorial composition of almost all of these subtests is described successfully by the CHC taxonomy, regardless of whether or not they were designed to tap into CHC abilities. The invariant CHC broad ability factors provide additional support for the CHC-based cross-battery assessment approach, particularly with regard to its guidelines for combining subtests from different batteries to create CHC composites (Flanagan et al., 2012).

It should also be noted that the findings concerning CHC theory, in general, have important, albeit indirect, implications for assessment practice. Many psychologists use CHC theory as a guide in interpreting intelligence test results. These findings suggest that the underlying theory is indeed valid, and therefore, interpretations based on that theory are more likely to be valid. If, in contrast, the underlying theory were flawed, then any interpretations based on that theory would have a much lower chance of being valid. Valid interpretations are much more likely to yield useful conclusions and recommendation than are invalid ones.

[7] Sample sizes are different and, in general, standardized effects are not interpreted for dummy variables.

Likewise, consider if invariance of factors across tests was untenable. The interpretation of a Gc or a Gv factor from one test would mean something different from the interpretation of a Gc or Gv factor from a different test battery. These results suggest that the factors (at least the ones tested here) represent the same underlying constructs whether evaluated on Test A or Test B, in combination, or alone. Interpretations based on CHC theory may or may not be valid and useful, but they stand a much higher chance of being demonstrably so if the underlying constructs are stable across measures. In sum, these findings primarily have implications for theory, but theory, in turn, has important implications for practice.

4.2. Factorial composition of subtests

The information in Table 5 should provide useful information about the factorial composition of the subtests used in this research. It is clear that *g*, broad abilities, and something specific are measured in each subtest. For the most part both *g* and broad ability effects (or loadings), independent of *g*, on subtests are moderate to strong. Gf subtests, on average, had the strongest loadings on *g*. Gc subtest loadings on *g* were stronger than the Gf loadings on *g* when heterogeneity was not controlled, but they were weaker when it was controlled. Gc loadings on *g* have likely been overestimated in previous research due to population heterogeneity and the current findings provide further evidence (see Ashton & Lee, 2006) against a Gc interpretation of *g* that has been proposed (Gignac, 2006).

This research did not include statistical tests of every plausible rival hypothesis with regard to individual subtest classification, but there were some subtest-specific findings that should be of interest to those who use intellectual assessment in practice and research. First, the WJ III Picture Recognition subtest appeared to measure Associative Memory, which is a more narrow measure of Glr, rather than Gv. Researchers have recently associated this subtest with a Glr factor (Kaufman et al., 2012), and it may be argued that it measures Associative Memory (or initial learning efficiency) because an examinee is required to locate objects that were exposed previously for 5 seconds. Although it does not contain multiple learning trails and objects are not paired with different information initially, perhaps the information that is required to be stored is more than can be maintained in short-term memory, and subsequently the information is accessed when it is either paired with the same objects or differentiated from other objects. Alternatively, Carroll (1993) used the term Gy to refer to a general memory and learning factor. Both Associative Memory and Visual Memory were indicators of this Gy factor. That is, perhaps these are indicators of a broader memory and learning factor. In light of this finding, however, it should be noted that the Picture Recognition specific variance correlated with the WJ III Spatial Relations subtest specific variance. Scores from these two subtests are combined on the WJ III to obtain a Gv composite score. This correlated specific indicates that these two subtests do share some variance, and it should be noted that the residual variances for these two subtests were not invariant when compared to estimates obtained from the norming sample. Future research may further clarify what Picture Recognition or similar tasks measure. All of the other WJ III subtests loaded on the expected CHC factors.

There were two subtest-specific findings related to the WISC-III subtests. Arithmetic from the WISC-III was found to measure Gsm, or more specifically working memory (cf. Keith & Witta, 1997; Kranzler, 1997; Phelps et al., 2005). The current study, however, did not have additional tests of quantitative reasoning or applied math, which precluded tests of alternative hypotheses. A second finding was that the supplemental subtest of Picture Completion was not a good measure of Gv, and instead measured Gc. On the surface, Picture Completion does not seem like a good measure of Gc, although examinees are generally required to articulate the name of the a missing part. The naming requirement, perhaps, is what may lead it to load on Gc. The subtest has loaded on Gc in other cross-battery research (Phelps et al., 2005), but it should also be noted that other than the KABC-II Gestalt Closure subtest, which was cross-loaded on Gc and Gv, Picture Completion had the lowest loading on Gc. In contrast, all WISC-IV subtests loaded on the expected factors, including the Matrix Reasoning and Picture Concepts subtests, which have been identified previously as indicators of Gf (Keith et al., 2006).

There were also a few interesting findings related to KABC-II supplemental subtests. First, Hand Movements was found to measure Gsm. This finding is consistent with the classification provided by the KABC-II authors (Kaufman & Kaufman, 2004), and clarifies results from a previous within-battery factor analysis of the KABC-II data (Reynolds et al., 2007). Alternatively, Gestalt Closure was not a good measure of any broad ability (see Table 5). It may measure *g*, although not strongly, plus something specific. Practitioners should keep this finding in mind while administering Gestalt Closure, as it may not be an optimal substitute for a Gv or Gc subtest. Although Gestalt Closure was not a strong measure of CHC broad abilities or *g*, it does not mean that what it measures is unimportant. Rather, this subtest may simply measure something more specific. Future research should consider what is being measured by this subtest because whatever it is, it seems to be mostly independent of common factors, including *g*. Some skills, for example recognition of human faces, may operate independently of common factors (Wilmer et al., 2010), and study of these specific skills offer an intriguing line of future research.

In all, the findings related to the factorial composition of the subtests were consistent with CHC classifications. Most departures were related either to supplemental tests or to subtests that have been found to migrate to different factors in other research (e.g., Phelps et al., 2005; Reynolds et al., 2007). For the most part, examiners should be comfortable with the CHC classifications of these subtests (e.g., Flanagan, Ortiz, & Alfonso, 2013).

4.3. Second-order g factor loadings

The pattern and magnitude of the second-order loadings, when population heterogeneity was controlled for, were amazingly similar to research in which CHC construct representation was introduced into automatic item generation (Arendasy et al., 2008). The standardized loadings of Gc (.75) and Gf (.98) were exactly the same, and Gsm (.75) and Gv (.76) loadings were within .01. The findings are notable because the second-order loadings (i.e., structural coefficients of latent constructs) are three times

removed from the item level, and considerable differences in item development (automatic generation versus generation by different test development teams) did not alter the structural or substantive parameters related to the latent constructs. These findings strongly support CHC theory as a framework for understanding psychometric intelligence.

Heterogeneity related to SES and sex influenced the second-order factor loadings; thus, homogeneity of the sample should be considered when interpreting factor loadings. The covariates together explained about 10% of *g* variance, and the biggest change when controlling for heterogeneity was the Gc factor loading on *g*, although Gsm and Gv factor loadings decreased as well. In general, however, Gc factor loadings on *g* may be overestimated in other studies. That is, SES and sex are a common influence on Gc and *g* (and to a lesser extent Gsm and Gv), whereas with the other factors SES and sex affects them primarily through *g*.

The second largest *g* loading was from the Associative Memory factor, which is interesting because these tasks to some extent represent new learning. If a test is administered by a well-trained examiner, one who eliminates extraneous influences and keeps a child motivated and interested, then the examinee would be likely to invest their *g* ("historical Gf") into tasks that require new learning (Cattell, 1987). It has been hypothesized that if scores from several new learning tasks were factor analyzed, the general factor obtained from that analysis would be highly correlated with a *g* factor extracted from an intelligence battery (Jensen, 1989). Our finding seems to provide some support this hypothesis. It would be interesting to investigate if this loading would increase if scores from additional and more varied new learning tasks were included.

Gf had the strongest loading on *g*, demonstrating a perfect correlation with the second-order *g* factor. The finding has substantial theoretical implications regarding intelligence. First, although a very strong relation between *g* and Gf is commonly found in research, a perfect statistical relation between Gf and *g* is not ubiquitous (cf. Bickley et al., 1995; Carroll, 2003; Gustafsson, 1984; Kvist & Gustafsson, 2008; Undheim & Gustafsson, 1987). One popular explanation for the different findings across studies is that there are too few measures of Gf, and a formal test of Gf's linear independence from *g* has been difficult without an adequate sampling of measures (Carroll, 2003). The current study addressed this limitation because seven measures of Gf were available, and only two of the seven measures were matrix reasoning type tasks, thus the indicators varied in content. A second noted limitation of prior research has been a lack of statistical power to detect a difference from a perfect relation (Matzke, Dolan, & Molenaar, 2010). Here, the hypothesis of a perfect relation between the two was tested formally within a CFA model. The Gf loading on *g* was very large in magnitude (.98) when population heterogeneity was controlled for in the other common factors, so it may be argued that practically speaking, power was not overly important. Moreover, Associative Memory had the second highest loading on *g*. When the Associative Memory residual was constrained to zero, model fit degraded substantially $\Delta\chi^2\,(1) = 47.70, p < .001$. Associative Memory was independent from *g*, Gf was not.

Another explanation for inconsistent perfect Gf–*g* relation findings across studies was raised by Kvist and Gustafsson (2008) who suggested that a less-than-perfect relation may be a result of population heterogeneity. They found that only when populations were homogeneous in their opportunity to learn was the Gf–*g* relation perfect. Drawing from Cattell's investment theory, Kvist and Gustafsson hypothesized that historical Gf is a causal factor in all learning, especially early on; however, Gf and "*g*" differentiate once differences in opportunity to learn arise. In the current study, heterogeneity induced by SES and sex was controlled. Thus, the findings from the current study are partially consistent with those from Kvist and Gustafsson in that SES differences in *g* were controlled (i.e., participants were statistically equalized on SES). Nevertheless, even without the SES covariate Gf was not linearly independent of *g*, although the magnitude of the loadings was slightly less. Population heterogeneity related to sex and SES (which may or may not be considered a proxy of opportunity to learn) was either not important enough to produce differentiation in Gf and *g*, or there was not much variability in opportunity to learn in these samples, or in the United States at large, as related to SES.

A perfect relation between Gf and *g* is of theoretical import. Some researchers have suggested that Cattell and Horn's description of Gf is similar to that used by Spearman to describe *g* (Gustafsson, 1984). Moreover, if there is a perfect relation between the two, then *g* may be identified invariantly by Gf tests in the same way that Gf factors are identified invariantly by Gf tests. Such an argument has been used to counter the argument by researchers who have contended that *g* lacks factorial invariance. Alternatively, other researchers have argued that if Gf and *g* correlate perfectly then the second-order factor is not a "general factor" at all because the first-order Gf factor is independent of the first-order factors, and it does not account for their intercorrelations (Horn & McArdle, 2007). Last, other researchers have noted that if the two factors are identical, then it does not make sense to describe a second-order *g* factor that is ill-defined when it is the same as a well-defined Gf first-order factor (Matzke et al., 2010). These are all interesting arguments, and the question of whether *g* is redundant or Gf is redundant is one of the important unresolved questions in intelligence research, especially with regard to the status of *g* as an underlying causal variable (Gustafsson, 1984; Horn & McArdle, 2007; Jensen, 1998).

Although the Gf–*g* relation is of theoretical import, it may also have practical implications for research and assessment. For example, if a quick estimate of *g* is needed for research or assessment, it may be best to sample from the Gf domain. Future research, however, should test how well such short Gf-based tests compare to more traditional brief measures of intelligence, ones that usually sample from several broad abilities.

4.4. Limitations and recommendations for research

The findings of this research should be interpreted within the context of the limitations of the study. First, despite different theories of intelligence (see Flanagan & Harrison, 2012), we did not test the CHC model against any competing theories. For example, the verbal, perceptual, and image rotation (VPR) model is a higher-order model that departs somewhat from the CHC taxonomy (Johnson & Bouchard, 2005, also see Vernon, 1965). Future research should test the CHC taxonomy against the VPR

model and other theories of intelligence. Use of a reference-variable approach to maximize the breadth of constructs measured in any given sample would be an excellent consideration for designing future research to test plausible rival hypotheses.

Second, this study included subtests that were only expected to target four broad abilities and one narrow ability. Additional broad abilities should be included in future research. For example, there are some questions with regard to the Glr factor (see Carroll, 1993; Keith & Reynolds, 2010; Schneider & McGrew, 2012). Or based on findings from this study, Carroll's G*y* factor, or some variation of that factor, needs to be investigated in more depth. In addition, some of the more narrow abilities could have modeled more explicitly in this research. For example, we correlated the Memory Span residuals rather than explicitly modeling this Memory Span factor. More refined classification of CHC narrow abilities is also the next logical next step for future research.

Future research should investigate the best designs for planned missing research. For example, although there were no Gs subtests in the KABC-II, there were Gs subtests included with the WISC and WJ III data. We ran some additional analyses, although not reported in this study, in which these Gs subtests were included in CB-CFAs and assigned to a Gs factor. They had substantial loadings on this factor, and the overall model fit well. That is, although there were not "reference" tests of processing speed in the KABC-II battery, a Gs factor was tenable using subtests from the other batteries. Such findings should be studied in more detail in future research because these types of research designs may be used to improve data collection.

Next, although the purpose of this study was not to study the Flynn effect (Flynn, 2007; Flynn & Weiss, 2007), some aspects of the Flynn effect, which is generally thought to be related to *g* (i.e., IQs are increasing), were supported. Different groups of children took the WISC-III and WISC-IV, but scores from the WISC-IV (which has a more recent norm sample) were generally lower. There was a small, yet statistically significant Flynn effect related to *g*, but there were also differences due to a specific mean difference on the Similarities subtest, which is inconsistent with a Flynn effect (increase in a specific ability related to Similarities beyond *g*). Although our WISC-IV sample size was small, the finding is consistent in that Similarities often shows the largest change over time. There are few studies that have studied the Flynn effect as related to measurement invariance (Wicherts et al., 2004). Measurement invariance is important in determining whether change in mean scores is due to *g* or specific factors. Here, the largest change was due to a specific factor related to the Similarities subtest. Interpretations of possible reasons for these differences are discussed elsewhere (cf. Flynn & Weiss, 2007; Kaufman, 2010), but studying whether the Flynn effect is mostly due to changes in *g* versus specific factors should be considered in future research.

Last, more theorizing is needed as to why a perfect correlation between Gf and *g* exists. The investment hypothesis has been proposed as one explanation (Kvist & Gustafsson, 2008). Theories should also integrate empirical findings related to Spearman's law of diminishing returns in this research (Detterman & Daniel, 1989; Reynolds, Hajovsky, Niileksela, & Keith, 2011). In this empirically observed phenomenon, correlations between subtests (or factors) decrease as a function of *g*. One interesting question that should be investigated is whether Gf and *g* differentiate (i.e., *g*–Gf correlation becomes less than one) at higher levels of *g*.

4.5. Summary

CHC broad ability factors were invariant across selections of populations and variables. The factorial composition of an overwhelming majority of subtests from several individually administered intelligence test batteries was easily predictable from CHC theory and CHC-based prior research when all of the subtests from the batteries were analyzed simultaneously and when different samples of people were administered the tests. Gf was statistically indistinguishable from *g*, and statistically controlling for SES and sex resulted in lower Gc loadings. Last, it was shown certain planned missingness designs, such as those that include reference variables, allow for a greater breadth of broad ability factor indicators to be used while studying intelligence.

References

Arendasy, M. E., Hergovich, A., & Sommer, M. (2008). Investigating the g saturation of various stratum-two factors using automatic item generation. *Intelligence*, *36*, 574–583.

Ashton, M. C., & Lee, K. (2006). "Minimally biased" *g* loadings of crystallized and non-crystallized abilities. *Intelligence*, *34*, 469–477.

Baker, T. J., & Bichsel, J. (2006). Personality predictors of intelligence: Differences between young and cognitively healthy older adults. *Personality and Individual Differences*, *41*, 861–871.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246.

Bickley, P. G., Keith, T. Z., & Wolfle, L. M. (1995). The three-stratum theory of cognitive abilities: Test of the structure of intelligence across the life span. *Intelligence*, *20*, 309–328.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York, NY: Cambridge University Press.

Carroll, J. B. (1996). A three-stratum theory of intelligence: Spearman's contribution. In I. Dennis, & P. Tapfield (Eds.), *Human abilities. Their nature and measurement* (pp. 1–18). Mahwah, NJ: Erlbaum.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5–21). Boston, MA: Pergamon.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action.* Amsterdam, The Netherlands: Elsevier.

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

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.

Darlington, R. B. (1990). *Regression and linear models.* New York, NY: McGraw-Hill.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis–Kaplan Executive Function System.* San Antonio, TX: Pearson.

Detterman, D. K., & Daniel, M. H. (1989). Correlations of mental tests with each other and with cognitive variables are highest for low IQ groups. *Intelligence*, *13*, 349–359.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197.

Flanagan, D. P., Alfonso, V. C., & Ortiz, S. O. (2012). The cross-battery assessment (XBA) approach: An overview, historical perspective, and current directions. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 459–483) (3rd ed.). New York, NY: Guilford Press.
Flanagan, D. P., & Harrison, P. L. (Eds.). (2012). *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed.). New York, NY: Guilford Press.
Flanagan, D. P., & McGrew, K. S. (1998). Interpreting intelligence tests from modern Gf–Gc theory: Joint confirmatory factor analysis of the WJ-R and Kaufman Adolescent and Adult Intelligence Test (KAIT). *Journal of School Psychology*, *36*, 151–182.
Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). *Essentials of cross-battery assessment* (3rd ed.). Hoboken, NJ: Wiley.
Floyd, R. G., Bergeron, R., Hamilton, G., Parra, G. R., & McGrew, K. S. (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.
Floyd, R. G., Shands, E. I., Rafael, F. A., Bergeron, R., & McGrew, K. S. (2009). The dependability of general-factor loadings: The effects of factor-extraction methods, test battery composition, test battery size, and their interactions. *Intelligence*, *37*, 453–465.
Flynn, J. R. (2007). *What is intelligence?* New York, NY: Cambridge University Press.
Flynn, J. R., & Weiss, L. G. (2007). American IQ gains from 1932–2002: The WISC subtests and educational progress. *International Journal of Testing*, *7*, 209–224.
French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, *13*, 378–402. S.
Gignac, G. (2006). Evaluating subtest 'g' saturation levels via the single-trait correlated uniqueness (STCU) SEM approach: Evidence in favor of crystallized subtests as the best indicators of 'g'. *Intelligence*, *34*, 29–46.
Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Planned missing-data designs in analysis of change. In L. M. Collins, & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 335–353). Washington, DC: American Psychological Association.
Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*, 323–343.
Gustafsson, J. -E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, *8*, 179–203.
Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM fit indexes with respect to violations of uncorrelated errors. *Structural Equation Modeling*, *18*, 36–50.
Hoogland, J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods and Research*, *26*, 329–367.
Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, & R. W. Woodcock (Eds.), *WJ-R technical manual* (pp. 197–232). Itasca, IL: Riverside Publishing.
Horn, J. L., & McArdle, J. J. (2007). Understanding human intelligence since Spearman. In R. Cudeck, & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 205–247). Mahwah, NJ: Erlbaum.
Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf–Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (pp. 53–91). New York, NY: Guilford Press.
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. (1989). The relationship between learning and intelligence. *Learning and Individual Differences*, *1*, 37–62.
Jensen, A. R. (1998). *The g factor.* Westport, CT: Prager.
Jensen, A. R., & Weng, L. -J. (1994). What is a good *g*? *Intelligence*, *18*, 231–258.
Johnson, W., & Bouchard, T. J. (2005). The structure of human 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., Kievit, R. A., Dolan, C., & van der Maas, H. (2011). On the interpretation of the CHC factor Gc. *Intelligence*, *39*, 292–302.
Kaufman, A. S. (2010). "In what way are apples and oranges alike?" A critique of Flynn's interpretation of the Flynn Effect. *Journal of Psychoeducational Assessment*, *28*, 382–398.
Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children—2nd ed.* Circle Pines, MN: AGS.
Kaufman, S. B., Reynolds, M. R., Liu, X., Kaufman, A. S., & McGrew, K. S. (2012). Are cognitive g and academic achievement g one and the same g? An exploration on the Woodcock-Johnson and Kaufman tests. *Intelligence*, *40*, 123–138.
Keith, T. Z. (1999). Effects of general and specific abilities on student achievement: Similarities and differences across ethnic groups. *School Psychology Quarterly*, *14*, 239–262.
Keith, T. Z. (2005). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 581–614) (2nd ed.). New York, NY: Guilford Press.
Keith, T. Z. (2006). *Multiple regression and beyond.* Boston, MA: Pearson.
Keith, T. Z., Fine, J. G., Taub, G. E., Reynolds, M. R., & Kranzler, J. H. (2006). Higher order, multisample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children—4th ed.: What does it measure? *School Psychology Review*, *35*, 108–127.
Keith, T. Z., Kranzler, J. H., & Flanagan, D. P. (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 ed.). *School Psychology Review*, *30*, 89–119.
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*, 635–650.
Keith, T. Z., & Reynolds, M. R. (2012). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 758–799) (3rd ed.). New York, NY: Guilford Press.
Keith, T. Z., & Witta, E. L. (1997). Hierarchical and cross-age confirmatory factor analysis of the WISC-III: What does it measure? *School Psychology Quarterly*, *12*, 89–107.
Kranzler, J. H. (1997). What does the WISC-III measure? Comments on the relationship between intelligence, working memory capacity, and information processing speed and efficiency. *School Psychology Quarterly*, *12*, 110–116.
Kvist, A. V., & Gustafsson, J. -E. (2008). The relation between fluid intelligence and the general factor as a function of cultural background: A test of Cattell's investment theory. *Intelligence*, *36*, 422–436.
Major, J. T., Johnson, W., & Bouchard, T. J. (2011). The dependability of the general factor of intelligence: Why small, single-factor models do not adequately represent g. *Intelligence*, *39*, 418–433.
Markwardt, F. C. (1998). *Peabody Individual Achievement Test — Revised Normative Update.* Circle Pines, MN: AGS.
Matzke, D., Dolan, C., & Molenaar, D. (2010). The issue of power in the identification of "g" with lower-order factors. *Intelligence*, *38*, 336–344.
McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research*, *29*, 409–454.
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*, 525–543.
Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*(11 Suppl. 3), S69–S77.
Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd ed.). Boca Raton, FL: Taylor & Francis.
Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.
Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.
Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 599–620.
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*, 66–88.

Reynolds, M. R., Hajovsky, D. B., Niileksela, C., & Keith, T. Z. (2011). Spearman's Law of Diminishing Returns and the DAS-II: Do *g* effects on subtest scores depend on *g*? *School Psychology Quarterly*, *26*, 275–289.
Reynolds, M. R., Keith, T. Z., Fine, J. G., Fisher, M. F., & Low, J. A. (2007). Confirmatory factor analysis of the Kaufman Assessment Battery for Children—2nd ed.: Consistency with Cattell–Horn–Carroll theory. *School Psychology Quarterly*, *22*, 511–539.
Reynolds, M. R., Keith, T. Z., Ridley, K. P., & Patel, P. G. (2008). Sex differences in latent general and broad cognitive abilities for children and youth: Evidence from higher-order MG-MACS and MIMIC models. *Intelligence*, *36*, 236–260.
Roid, G. H. (2003). *Stanford-Binet Intelligence Scales, 5th ed.: Technical manual.* Itasca, IL: Riverside Publishing.
Sanders, S., McIntosh, D. E., Dunham, M., Rothlisberg, B. A., & Finch, H. (2007). Joint confirmatory factor analysis of the Differential Ability Scales and the Woodcock-Johnson Tests of Cognitive Abilities—3rd ed. *Psychology in the Schools*, *44*, 119–138.
Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.
Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53–61.
Schmitt, N., Golubovich, J., & Leong, F. T. L. (2011). Impact of measurement invariance on construct correlations, mean differences, and relations with external correlates: An illustrative example using Big Five and RIASEC measures. *Assessment*, *18*, 412–427.
Schneider, J. W., & McGrew, K. S. (2012). The Cattell–Horn–Carroll model of intelligence. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 99–144) (3rd ed.). New York, NY: Guilford Press.
Schrank, F. A., Miller, D. C., Wendling, B. J., & Woodcock, R. W. (2010). *Essentials of WJ III cognitive abilities assessment* (2nd ed.). Hoboken, NJ: Wiley.
Steiger, J. H. (1998). A note on multiple sample extensions of the RMSEA fit index. *Structural Equation Modeling*, *5*, 411–419.
Steiger, J. H., & Lind, J. (1980). Statistically based tests for the number of common factors. *Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA*.
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 III. *School Psychology Quarterly*, *19*, 72.
Thorndike, R. L. (1987). Stability of factor loadings. *Personality and Individual Differences*, *8*, 585–586.
Thurstone, L. L. (1947). *Multiple-factor analysis: A development and expansion of the vectors of the mind.* Chicago, IL: University of Chicago Press.
Tucker, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika*, *23*, 111–136.
Undheim, J. O., & Gustafsson, J. -E. (1987). The hierarchical organization of cognitive abilities: Restoring general intelligence through the use of linear structural relations (LISREL). *Multivariate Behavioral Research*, *22*, 149–171.
van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*, 842–861.
Vernon, P. E. (1965). Ability factors and environmental influences. *American Psychologist*, *20*, 723–733.
Wechsler, D. (1991). *Wechsler Intelligence Scale for Children—3rd ed.: Manual.* San Antonio, TX: Pearson.
Wechsler, D. (2003). *The WISC-IV—Technical manual.* San Antonio, TX: Pearson.
Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C., Boomsma, D. I., et al. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, *32*, 509–537.
Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., et al. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 5238–5241, http://dx.doi.org/10.1073/pnas.0913053107.
Woodcock, R. W. (1990). Theoretical foundations of the WJ-R measures of cognitive ability. *Journal of Psychoeducational Assessment*, *8*, 231–258.
Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock–Johnson Psychoeducational Battery-Revised: Tests of Cognitive Ability.* Chicago, IL: Riverside Publishing.
Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III Tests of Cognitive Abilities.* Itasca, IL: Riverside Publishing.
Wothke, W. (2000). Longitudinal and multi-group modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple group data: Practical issues, applied approaches, and specific examples* (pp. 219–240). Hillsdale, NJ: Erlbaum.