Please use the following citation when referencing this work:

Dombrowski, S. C., Beaujean, A. A., McGill, R. J., & Benson, N. F. (2018). The Woodcock-Johnson IV Tests of Achievement provides too many scores for clinical interpretation. *Journal of Psychoeducational Assessment*. Advance online publication. doi: 10.1177/0734282918800745

The Woodcock-Johnson IV Tests of Achievement Provides Too Many Scores for Clinical Interpretation

Stefan C. Dombrowski

Rider University

A. Alexander Beaujean

Baylor University

Ryan J. McGill

William & Mary

Nicholas F. Benson

Baylor University

Author notes

Correspondence concerning this article should be addressed to Stefan C. Dombrowski, Professor & Director, School Psychology Program, Department of Graduate Education, Leadership & Counseling, Rider University, 2083 Lawrenceville Road, Lawrenceville, NJ 08648, USA. Email: sdombrowski@rider.edu

Abstract

The Woodcock-Johnson Tests of Achievement, Fourth Edition (WJ IV ACH; Schrank, Mather & McGrew, 2014) is purported to align with CHC Theory and offers upwards of 20 scores within its interpretive and scoring system. The Technical Manual does not furnish validity evidence for the scores reported by the scoring system, suggesting that evidentiary support may be incomplete. Exploratory bifactor analysis (maximum likelihood extraction with a bigeomin [orthogonal] rotation) was applied to the two school-aged correlation matrices at age 9-19. Results indicated non-alignment with CHC theory and do not support the interpretation of most of the scores suggested by the scoring system. Instead, the results of this study suggest that the loading patterns diverge significantly from the interpretive system produced by the WJ IV ACH. Only the academic fluency and academic knowledge clusters emerged following the use of EBFA. Implications for clinical interpretation of the WJ IV ACH are offered.

Keywords: Woodcock-Johnson IV Achievement, exploratory bifactor analysis, factor analysis, measurement, clinical interpretation

The Woodcock-Johnson IV Tests of Achievement Provides Too Many Scores for Clinical Interpretation

A basic tenant of measurement is that symbolic values (e.g., numbers) need to represent attributes (Mari, Carbone, & Petri, 2015). One can, of course, create numerical indices that represent mixtures of multiple possible attributes, but this is not really measurement—or at least the type of measurement normally desired by psychologists—since the attributes are defined by the instrument (Fayers & Hand, 2002). Thus, the fundamental aspect of validity for scores from a psychological instrument is that the numbers provide accurate information about a well-defined attribute (Krause, 2012).

Unfortunately, the notion that scores need to represent specific, well-defined attributes is not currently common practice when creating commercially available psychological instruments (e.g., Krause, 2005). Instead, the more typical approach is for test publishers/authors to provide a plethora of scores for a given psychological instrument which adhere to a variety of (perhaps incompatible) theories, but not furnish evidence that the instrument measures all of the attributes it claims to measure (Author & Author, in press). An example of such an instrument is the Woodcock-Johnson Tests of Achievement, Fourth Edition (WJ IV ACH; Schrank, Mather & McGrew, 2014).

The WJ IV ACH is one of the more popular nationally-normed individually-administered tests of academic achievement. It is cast within the Cattell-Horn-Carroll (CHC) theoretical framework (McGrew, 2005; Schneider & McGrew, 2012) and the Technical Manual indicates that the WJ IV ACH measures several CHC attributes, including Reading/Writing (Grw), Quantitative Reasoning (Gq), Academic Knowledge/Crystallized Ability (Gc), and Academic Fluency/Processing Speed (Gs) (McGrew, LaForte, & Schrank, 2014). A closer inspection of the WJ IV ACH scores, however, reveals that not all the scores available for interpretation stem from CHC theory, and some of the scores appear to be unrelated to any defined attribute.

The scoring system for the WJ IV ACH produces two total achievement scores (Brief and Broad Achievement) and 19 additional cluster scores (e.g., broad reading, basic reading, math problem solving, basic writing skills). The scores are shown in Table 1. Validity evidence for the scores is largely correlational: examination of correlations with scores from other academic achievement instruments as well as loadings of the tests on achievement-related latent variables. Oddly, the latent variable model's fit included not only tests from the WJ IV ACH but also tests from the Cognitive and Oral Language instruments. Consequently, the achievement-related latent variables were partially defined by non-achievement tests and the achievement-related tests loaded on non-achievement latent variables. This is appropriate for furnishing validity evidence for CHC theory but it may be questioned if attempting to establish validity for the WJ IV ACH in isolation. Further confounding the attempt to establish structural validity evidence, three recent exploratory investigations suggested that both the WJ IV Cognitive and full test battery were complexity determined and struggled to align with CHC theoretical structure and scoring system posited in the technical manual (Dombrowski, McGill & Canivez, 2017; 2018a; 2018b). Moreover, there was no discussion of the structural relationship among tests and cluster scores. Thus, it is difficult to understand precisely what attributes all the WJ IV ACH scores represent

The undetermined nature of what the WJ IV ACH scores represent is very problematic (Borsboom, Mellenbergh, & van Heerden, 2004). Given the general lack of technical definitions about the attributes these scores represent (and justification that those attributes are represented by the available scores) coupled with an insufficient investigation (or at least reporting) of the relation among the tests and their respective factor/cluster scores, many questions still remain about the instrument's scores. For example, what do the numbers actually represent? Does it make sense to interpret all the scores that are provided to users?

The lack of understanding of the structure of individually administered tests of academic achievement is not a concern solely relegated to the WJ IV ACH. It is a problem for the entire field of clinical assessment particularly academic achievement. A survey of the extant individually-administered academic achievement validity literature indicates that there is a dearth of relevant structural validity investigations. Beyond the achievement test’s technical manuals, there are only four investigations of the factor structure of individually administered tests of academic achievement, whether broad band or narrow band. Users, therefore, are left to rely upon the structure presented in various academic achievement test’s technical manuals.

Reynolds (1979) investigated the Peabody Individual Achievement Test (PIAT) and found that its structure had two factors consistent with Cattell’s concept of fluid and crystallized intelligence. Williams and Eaves (2001) investigated the Woodcock Reading Mastery Test, a narrow band measure of reading, and found that the instrument’s structure was unidimensional (broad reading) and not multidimensional as indicated in the instrument’s Technical Manual. Williams, Fall, Eaves, Darch, and Woods-Groves (2007) also explored the structure of the KeyMath Normative Update and provided evidence for a singular global math factor rather than the three-factor solution posited by the Technical Manual. More recently, Dombrowski (2015) applied the Schmid-Leiman (1957) orthogonalization procedure to the core and extended battery of the Woodcock-Johnson Tests of Achievement, Third Edition (WJ III ACH; Woodcock, McGrew, & Mather, 2001) across the 9-13 and 14-19 age ranges. Dombrowski’s results suggested the prominence of a general achievement factor and a different number and composition of group factors than what was posited in the manual. Dombrowski failed to locate distinct oral language, reading, and writing factors but found evidence for standalone mathematics (Gq) and academic fluency (Gs) factors. As a result, Dombrowski suggested caution when moving to interpretation of specific group factors because of their complexity in several cases. Further supporting this position, omega estimates for the group factors produced by Dombrowski's study were not sufficiently high for individual interpretation, calling into question their potential clinical utility. With the exception of an academic fluency factor, Dombrowski's findings did not show evidence for the numerous cluster scores reported by the WJ III scoring system.

Consequently, Dombrowski (2015) called for additional research into the theoretical structure of academic achievement tests, noting that they are widely used but infrequently subjected to the same type of validity evaluation as other types of psychological tests (e.g., cognitive ability, personality). As an example, the technical manual for the most recent version of the widely utilized Wechsler Individual Achievement Test (WIAT-III; Pearson, 2009) just provides correlations with other tests as evidence of validly—a common practice but one that has been regarded as insufficient by methodologists (i.e., Borsboom, Mellenbergh, & van Heerden, 2004).

Although the structure of the WJ IV ACH was analyzed as part of a conjoint investigation using tests from the Cognitive and Oral Language batteries, its theoretical structure was not independently examined using factor analysis nor were the amalgam of scores (~20) produced by the WJ IV scoring system. This omission is notable as many clinicians elect to administer the tests of achievement for diagnosis and classification decisions. As factor analytic studies provide the empirical basis for the scores that are developed for psychoeducational instruments such as the WJ IV ACH (Bandalos & Gerstner, 2016), validation of structure is critical given the role that achievement tests can play in important educational decisions such as special education eligibility determination and intervention planning.

The purpose of the current study was to investigate the factor structure of the WJ IV ACH independently from the subtests in the other two WJ IV instruments (i.e., oral language and cognitive). Thus, this study provided information needed to assist in determining the interpretive relevance of the scores that are provided to users of the measurement instrument. It also can aid in improving the theoretical understanding of the structure of one of the major academic achievement instruments used by clinicians and researchers.

Method

Sample and Instrument

The 9-to-13- and 14-to-19-year-old correlation matrices for the WJ IV ACH were ascertained from the instrument's Technical Manual (McGrew et al., 2014, p. 311-312). The matrices contained correlations for all 20 achievement subtests. Raw data were unavailable. The WJ IV Technical Manual described a planned missingness procedure for individuals who were not administered certain subtests where their normative scores were created using multiple imputation. The Technical Manual does not delineate how many participants had missing data on certain variables. As noted by Canivez (2017), extensive missing data may have impacted the WJ IV correlation matrices.

Data Analysis

Because the WJ IV ACH is structured hierarchically with an aggregate score (total achievement) and numerous more specific scores (e.g., cluster scores) that are intended for interpretation, exploratory bifactor analysis (EBFA; Jennrich & Bentler, 2011, 2012) may be a worthwhile analytical procedure. Some researchers argue that a bifactor conceptualization of academic achievement is also consistent with how Carroll how conceptualized his three stratum theory (Beaujean, 2015). EBFA permits the simultaneous assignment of variance to general and group factors yielding results that allow the calculation of metrics of clinical interpretative relevance (e.g., omega coefficients, explained common variance, subtest specificity).

Before conducting the EBFA, we examined different indexes related to the number of factors to extract. This included the Kaiser criterion (Kaiser, 1974), the visual scree test (Cattell, 1966), the minimum average partial test (MAP; Velicer, 1976), parallel analysis (factors and components at the 50th and 95th percentile; Horn, 1965), and the empirical Bayesian Information Criterion (BIC). Theoretical consistency and adherence with simple structure was also considered.

For the EBFA, we used maximum likelihood extraction and the bigeomin (orthogonal) rotation (Jennrich & Bentler, 2012). We analyzed the 9-13 and 14-19 years old correlation matrices separately. To aid in interpreting the results, we calculated explained common variance (ECV; Reise, Moore, & Haviland, 2010), omega coefficients (Brunner, Nagy, & Wilhelm, 2012; Reise, 2012) and H (Hancock & Mueller, 2001). These indices provide information concerning the relevance of the group factors and general achievement scores. Coefficient omega hierarchical, coefficient omega hierarchical subscale, and H were calculated to determine interpretive relevance of the general and group factors. Omega coefficients (i.e., omega hierarchical and omega hierarchical subscale) are construct-based reliability estimates that are more appropriate than coefficient alpha for multidimensional tests such as the WJ IV ACH. Reise (2012) and Reise, Bonifay and Haviland (2013) note that omega coefficients should exceed .50, but .75 is preferable to indicate sufficient construct based reliability for independent interpretation of a group or hierarchical factor. Hancock and Mueller's (2001) H was also used to determine emphasis for interpretive purposes. Rodriquez, Reise & Haviland (2016) indicate that it is difficult to specify group factors within a single instrument and it should only be done when H-values are higher than .70. We also calculated each subtest's specific reliable variance. Kaufman's (1975)[1] informal approach to subtest specific variance was referenced. Kaufman (1975) and Kaufman and Lichtenberger (2006) suggested that it is feasible to interpret a subtest in a measure of cognitive ability at the subtest level under two conditions: (1) when specific, reliable variance ≥ .25 and (2) when its specificity exceeds the estimate of variance that is attributable to measurement error.

Preliminary data analysis was conducted using the *psych* (Revelle, 2017) package in the R statistical programming language (R Development Core Team, 2017). The EBFA was conducted using Mplus Version 8 (Muthén & Muthén, 1998–2017).

Results

Table 2 displays factor extraction results for both age groups, suggesting the extraction of 3–6 factors. This is noticeably different from the 20+ scores the WJ IV ACH provides users to interpret. A six-factor solution produced factors that were most consistent with CHC theory. The results from this extraction are presented in Tables 3 and 4. The five factor extraction is presented in Online Supplement Tables A1 and A2. These results are presented as several indices recommended extraction of five factors. All tables contain factor loadings, explained common and total variance, communality estimates, uniqueness, subtest specificity, omega

[1] Kaufman (1975) and Kaufman and Lichtenberger (2006) cited Cohen (1959) to support the .25 threshold for interpretation of a cognitive ability subtest. Cohen never mentioned, nor implied, that a threshold of .25 was acceptable. Kaufman said that the threshold came from his own interpretation of Cohen's work coupled with David Wechsler's advice (A. Kaufman, personal communication, Feb 13, 2018).

coefficients, and H. Figures 1 and 2 provide a visual depiction of variance apportionment among general, group, and subtests across both age ranges.

Variance Partitioning Results

The use of EBFA permits the assignment of variance at different levels of generality (i.e., a general factor and group factors). For the six-factor solution (1 general and 5 group), across both age ranges, the general factor absorbed a considerable proportion of the explained common (.72 and .74) and total (.50 and .52) variance for ages 9–13 and 14–19, respectively. By contrast, the respective group factors assumed a substantially lower proportion of explained common (.04 to .09 and .03 to .07) and total variance (.03 to .06 and .02 to .05) for ages 9–13 and 14–19 (see Tables 3 and 4). Additionally, loadings on the general factor were high ranging from .57 to .86 (median=.70) at age 9–13 and .58 to .88 (median=.70) at age 14 –19. These results suggest the primacy of the general factor.

Omega Coefficients and H

Across both age ranges omega hierarchical and H exceeded the standard for confident clinical interpretation of the general factor. At ages 9–13 and 14–19, omega hierarchical was .90 and .93, respectively. H results (.96 for both age ranges) similarly indicated that the WJ IV ACH could be confidently interpreted at the general level. Omega hierarchical subscale, on the other hand, along with ECV and H suggested that there is insufficient group factor variance for confident clinical interpretation of the CHC-based indices. Tables 3 and 4 show that omega hierarchical subscale ranged from .13 to .33 and .13 to .32 for ages 9-13 and 14-19, respectively, well below the suggested threshold of .50 suggested for clinical interpretation. H was similarly low, with the highest H level of .63 for the age 9 to 13 Gc factor but below that level for all other group factors.

Subtest Specific Variance

Tables 3 and 4 show that subtest specificity ranged from .13 to .33 and .13 to .32 for ages 9-13 and 14-19, respectively, well below the suggested threshold of .25 suggested for clinical interpretation of individual subtests.

Pattern of Loadings

An investigation of the pattern of factor loadings indicates that the group factors appear to be related to five broad CHC-related attributes: Ga, Gq, Grw, Gs, and Gc. On the surface, this appears consistent with the theoretical perspective of the WJ IV; however, there are problems with superficial interpretation of these findings.

First, for both age groups the academic fluency subtests coalesced to form a Gs factor rather than load on their respective academic factors across both age ranges. The only exception was the Math Facts Fluency test which had a noticeable cross-loading on the Gq factor. Second, the Editing, Oral Reading, and Spelling tests did not saliently load on any group factor for either age range. However, these subtests all had substantial loadings on the general factor. Third, the most complexly determined factors were those produced from the reading and writing subtests. Some of the reading and writing tests fused together to form a Grw factor, while other tests (e.g., Letter-Word Identification, Spelling) that were designed to measure Grw (according to the Technical Manual) combined with other tests to cross-load what appears to be a Ga factor. This complex combination of subtests is problematic. Although numerous departures from simple structure (i.e., subtest cross-loading) were reported in the Manual, many of the scores that are provided to users are derived from weighted combinations of two or more subtests that fail to capture this dimensional complexity.

Discussion

In this study we investigated the structure of the WJ IV ACH. Ideally, the authors/publishers of the instrument would have explained the attributes the instrument was designed to measure. Then, the instrument's scores could be justified by explaining how each attribute's grammar (i.e., rules) was followed in creating scores for the attributes (Maraun, 1998). This was not done. Instead, the attributes the instrument was designed to measure are discussed independently of the scores provided for users to interpret. Subsequently, justification for the scores is implied through various correlation analyses—not of the WJ IV ACH tests, but of tests from all three of the WJ IV instruments together. Consequently, tests load on multiple factors, and the factors are defined by tests that are both theoretically consistent and inconsistent. Thus, clinicians may be left wondering what the scores represent and whether all of the more than 20 scores that are provided in the WJ IV scoring program should be interpreted. Relatedly, the results of Dombrowski's (2015) WJ III Achievement study suggest a similar lack of alignment with CHC theory and the underlying scores in the scoring program to that suggested in this study.

Although the face validity of academic achievement tests often makes it appear that the attributes the scores represent are "self-evident," they are not exempt from well-established validity requirements. In fact, the *joint standards* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) has a whole chapter devoted to tests used for educational assessment, including the need to fully understand an instrument's test structure and reliability prior to making high-stakes educational decisions. Nonetheless, clinical tests of academic achievement have only been sporadically studied in the peer-reviewed literature. Given the routine use of these tests in high stakes educational decisions (e.g., eligibility determination for special education and related services), this could well pose problems for the field.

Considering these issues, we re-analyzed the WJ IV ACH correlation data using EBFA, specifically investigating the consistency of the factor structure with the scores provided by the instrument to interpret. We used EBFA because the WJ IV ACH scores are hierarchical in nature—ranging from general to group. The results yielded a structure that was neither consistent with CHC theory nor the scoring system reported in the WJ IV Technical Manual. The major implication from this study is that it is unclear what attributes the WJ IV ACH scores represent.

The General Factor

Although results from EBFA indicate the presence of a general factor, it is unclear what attribute this factor represents. Is the WJ IV ACH primarily measuring some type of overarching general achievement attribute? If this is the case, then why was not such an attribute discussed in the technical manual? While the idea of such an attribute has been around for a long time (e.g., Cattell, 1963), it is ill-defined and we find it questionable that a major achievement instrument would look to measure such an attribute without devoting a substantial amount of text to it in the technical manual.

Is the general factor that was uncovered in this study consistent with the general factor found on tests of cognitive ability? Or, is it an academic achievement general factor? Whereas Dombrowski (2015) was agnostic to this question, Kaufman, Reynolds, Lui, Kaufman, & McGrew (2012) made the case for separate cognitive and academic general factors, suggesting that the two are distinct at the latent level. Nonetheless, Kaufman et al.'s (2012) study found a correlation between academic-*g* (ACH-*g*) and cognitive-*g* (COG-*g*) that averaged .83 (range .77 to 94), suggesting a fairly high level of isomorphism between the two constructs. In fact, this level of correlation is consistent with the correlation among the general factors that have been extracted from commercial tests of cognitive ability (Dombrowski, DiStefano & Noonan, 2004; Dombrowski & Noonan, 2004). As a result, it is yet unresolved whether the resulting general factor may simply be labeled as academic or cognitive.

Krause (2012) might contend that because only achievement tests were used, the general factor should be conceptualized as academic achievement. However, there is also the possibility that it reflects the higher order general cognitive ability factor found at the apex of CHC theory as many of the group factor dimensions located in the present study are consistent with that architecture. On the other hand, if psychometric *g* accounts for the majority of covariation among academic tasks then a general factor measured by academic achievement tests cannot also be the predominant cause of covariance. A general factor derived from academic achievement tests may not be isomorphic with psychometric *g* (i.e., Kaufman et al.'s conclusion) as not every general factor identified with factor analytic methods is psychometric *g*. While these factors have been found to correlate (Kaufman, Reynolds, Lui, Kaufman, & McGrew, 2012), the attribute theory for psychometric *g* purports that it be operationalized using a sufficient range of indicators which vary with respect to content, stimulus modality, and response modality (Jensen & Weng, 1993). Further, in accord with the attribute theory for *g*, indicators should include tasks that involve abstraction and novel problem-solving rather than relying solely on academic achievement tests involving acquired knowledge and skills. Given the absence of an attribute theory defining the general factor measured by academic achievement tests, it may be best to interpret this factor as a formative latent variable, or in other words a composite (Bolen & Bauldry, 2011). Regardless of whether the general factor is academic or cognitive, there is not a good theoretical conceptualization of what that factor represents. This line of thinking deserves further empirical attention.

Jensen (1993) noted that "A chief difference between the measurement of *g* and of achievement is that with tests of *g* our interest is mainly in the latent trait itself and not in the particular class of test items that reflect it and serve merely as vehicles for its measurement. In achievement testing, on the other hand, we are primarily interested in generalizing about the particular class of items in the achievement test. We want to know, for example, whether Johnny or Mary can add mixed fractions or do long division involving decimals" (p. 148). It is unclear what clinical relevance a global achievement score has, as it involves generalization across a very broad class of items.

When making educational diagnostic decisions, most clinicians tend to interpret academic group scores and subtests in accordance with the theoretical constructs postulated in federal guidelines for specific learning disabilities eligibility (i.e., basic reading, reading comprehension, math calculation, math reasoning, etc.). If a global composite score is available and test publishers recommend its interpretation, and the factor representing this score absorbs the majority of the reliable variance on these tests, then is this evidence to support the presence of a general factor of achievement? If one accepts the tenability of a general academic factor, then it appears that commercial ability tests such as the WJ IV ACH may only serve as indicators of global academic success. Could this finding be due to statistical artifact? In other words are there simply not enough measures of group factor abilities to differentiate them from general ability? If more or superior reading, writing or mathematics tests were added then would group factors dominate over the general factor? Although these questions cannot be answered without additional research, relative to cognitive tests, where indicators are often abstractions of latent abilities (e.g., Block Design) and thus the relationships to real world tasks are often illusive, the tasks on achievement tests bear a more direct relationship to the observable dimensions of achievement (e.g., reading, writing, and mathematics). Thus, scores reflecting well-specified academic attributes are likely to be readily interpretable, even if these scores are multidimensional (i.e., influenced by multiple factors).

Clusters

As previously mentioned, the WJ IV scoring software calculates myriad achievement cluster scores that were not modeled in the structural analyses reported in the Technical Manual. Only the academic fluency and academic knowledge clusters emerged following the use of EBFA in the present study. No support for the other cluster scores were found. Combined, these results suggest factorial complexity of the WJ IV ACH tests, which not only deviates from posited CHC theory but also deviates significantly from the scores reported by the WJ IV ACH scoring system.

Subtests

Although the preponderance of research emanating out of the cognitive ability literature has long recommended that subtest level analyses should be avoided (see McDermott, Fantuzzo, & Glutting, 1990; Watkins, 2003), other sources—primarily those that serve to guide clinician assessment practices but may not have been subjected to blind peer review (e.g., Groth-Marnat & Wright, 2016; Kaufman & Lichtenberger; 2006; Sattler, 2008)—have consistently encouraged interpretation of these tests at the subtest level, furnishing intuitive and logical support for this practice. The WJ IV Tests of Achievement Examiner's Manual (2014) does not expressly endorse the practice of subtest level interpretation but it does offer explicit details regarding the presumed abilities measured by the subtests. Consequently, reliable specific variance was incorporated in the present analyses. Using the criterion of .25 that Kaufman (1975) and Kaufman and Lichtenberger (2006) recommend as support for the interpretation of cognitive subtest scores, most of the WJ IV ACH have inadequate target construct variance for individual interpretation. At ages 9 to 13, only Spelling of Sounds, Applied Problems, Number Matrices, Writing Samples, Reading Recall, Humanities and Oral Reading exceed the recommended threshold for interpretation. At ages 14 to 19, only Spelling of Sounds, Number Matrices, Writing Samples, Reading Recall, and Oral Reading exceeded this threshold.

Validity Evidence Based on Test Content

As the interpretation of scores from tests of academic achievement often focus on generalizations about particular classes of items that reflect academic skills of interest, validity evidence based on test content seemingly is a more salient concern than is percent of target construct variance. Validity evidence based on test content, often referred to as construct validity, refers to "…the degree to which the content of a test is congruent with testing purposes" (Sireci & Baulker-Bond, 2014, p. 101). Notably, scant evidence of content validity is presented in the technical manual. Sireci and Faulkner-Bond (2014) recommend that congruence in ratings from multiple (i.e., 10 or more) subject matter experts be used as a criterion for determining if the content of an academic achievement test is sufficient with respect to relevance and quality, and that test alignment research be used to ensure that such tests are congruent with relevant curriculum frameworks and standards. It does not appear that these strategies were used when designing the WJ IV or when accumulating evidence to support the use and interpretation of its scores. Notably, overlap between content sampled in norm-referenced tests of achievement and local curricula has often been found to be very low (Shapiro, 2011).

While it is impractical for the publishers to study alignment with the local curriculum of all school districts that use the WJ IV ACH, there is an educational initiative (i.e., the Common Core State Standards Initiative; www.corestandards.org) that details essential English language arts, reading, and mathematical knowledge and skills that should be expected to be covered in the curricula of all U.S. K-12 schools. Alignment with the Common Core could be studied in an effort to better understand the extent to which score interpretations are supported by validity evidence based on test content.

Implications for Clinical Practice

The results of this study have important implications for the interpretation of broad band tests of academic achievement such as the WJ IV ACH. Despite producing a hierarchical total achievement score, academic achievement tests are frequently interpreted by focusing attention on the interpretation of group factor (i.e., index, composite) scores. As suggested by this study this interpretive approach must be undertaken cautiously for several reasons. First, subtest loadings suggest a degree of factorial complexity with several subtests experiencing salient cross-loadings on different factors. The exception includes Academic Fluency and Academic Knowledge. The academic fluency subtests formed a separate factor instead of contributing to their proposed academic factors. The Math Facts Fluency subtest was the sole exception. Also, two subtests at age 9-13 and three at age 14-19 load saliently on only the general factor without a salient group factor loading. The results of this study did not locate distinct reading, writing, and basic reading (and many other) factors furnished by the WJ IV scoring system. Even if the WJ IV ACH subtests perfectly loaded on their posited factors/scoring system clusters, we would still urge caution regarding the interpretation of many of the WJ IV ACH index level/cluster scores. There is simply insufficient unique variance to interpret most of these scores in isolation. Third, omega coefficients, ECV and H estimates indicate that users of the WJ IV ACH can interpret scores representing the general factor (i.e., Broad & Brief Achievement) with confidence, although additional interpretation of the scores related to group factors should be employed more cautiously, if at all. The general factor simply absorbs the majority of variance leaving insufficient residual variance for confident clinical interpretation of group factors. In totality, practitioners may do well to forgo using most of the cluster level scores calculated in the WJ scoring program (e.g., Basic Reading, Broad Reading, etc.) for high stakes educational decisions as it appears that many of these indices do not have sufficient structural support.

The results of this study also have implications that impact the interpretation of results from a series of predictive validity studies (e.g., Cormier, Bulut, McGrew & Frison, 2016; Cormier, Bulut, McGrew & Singh, 2017; Cormier, McGrew, Bulut & Fuanmoto, 2017). These studies demonstrated how selected indices from the WJ IV Cognitive may be used to predict academic achievement. However, if the underlying academic achievement constructs are poorly defined and lack structural validity evidence then the conclusions drawn from the aforementioned series of articles may be called to question. The constructs used in these predictive validity studies may not be as evident as the test publisher suggests.

Additionally, the results of this study have clinical implications that inform subtest interpretation. Examination of reliable specific variance that only 8 subtests at ages 9-13 and 7 subtests at ages 14-19 on the WJ IV ACH have sufficient specificity for isolated interpretation based on the criterion recommended for the interpretation of cognitive subtest scores. Please see

Figures 1 and 2 for a visual depiction of variance apportionment among general, group and subtests across both WJ IV ACH age ranges.

Admonitions against subtest interpretation are common in the empirical literature regarding tests of cognitive ability (Macmann & Barnett, 1997; Matarazzo & Prifitera, 1989; McDermott, Fantuzzo, & Glutting, 1990; Watkins, 2000, 2003; Watkins & Kush, 1994). Do these admonitions against subtest interpretation apply when interpreting academic achievement subtests—or at least the WJ IV ACH subtests? From a purely psychometric perspective, the answer is yes. However, as Jensen (1993) noted, the purpose of administering tests of academic achievement is typically to generalize about particular classes of items that reflect academic skills. For example, it makes little intuitive sense to create another spelling subtest so that a group factor can be created and a spelling index derived. A more practical solution is to expand the content of the spelling subtest to more generally reflect curricular standards, which in turn will permit a more reliable and valid evaluation of spelling abilities. Finally, provided extant WJ IV ACH subtests have sufficient validity evidence based on test content, interpreting these subtests may actually be a sounder practice than is interpreting the academic composites. While the WJ IV ACH composites are primarily measures of a general factor that appear to be poorly specified measures of academic attributes, WJ IV ACH subtests contain narrow classes of items that generally seem to reflect meaningful tasks and produce scores that are supported by estimates of reliability.

Study Limitations

Despite the consistency of these results across the school-age range for the WJ IV ACH, the present study is not without limitations. First, there is a legitimate question as to what model approach (i.e., oblique, bifactor, higher order) should be used to uncover the latent structure of academic achievement? We applied a bifactor model because the scores provided by the WJ IV are hierarchical in nature. But, there are other models to consider including a higher-order, oblique, and multi-unidimensional models. Additionally, limitations of the study include the use of correlation matrices instead of raw data and the replication on a sample presented in the technical manual. Analysis on independent samples will be worthwhile as will follow-up CFA studies comparing different conceptualization of academic achievement (i.e., oblique, higher order, bifactor).

Perhaps the most important limitation of the study is that we used correlational data to understand the attributes the instrument's scores represent. We are well aware of the irrelevance of using correlational data to define or validate attributes (Borsboom, Cramer, Kievit, Scholten, & Franić, 2009). Nonetheless, given the lack of conceptual clarity in the Technical Manual regarding what the sundry scores represent and the availability of the test correlations, we thought this was the best approach to take. Still, our results should be understood within the limitations of using empirical data to answer more conceptual questions (Jackson & Maraun, 1996).

Conclusion

In totality, the results of this study suggest caution regarding the interpretation of the WJ IV ACH scores. We base this statement on the lack of clear understanding of the attributes the scores represent, factorial complexity of the test scores (i.e., lack of alignment of subtests with their theoretically proposed factors or with the scoring system furnished by the WJ IV), and a lack of unique variance apportioned to the group factors and singular subtests to interpret all the scores that are provided to be interpreted. Given that there is an absence of evidence to verify what the instrument is really measuring, its scores should be used very cautiously in high-stakes decision making, or perhaps be limited to use in lower-stakes decisions such as academic screenings.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (4th ed.). Washington, DC: Authors.

Bandalos, D. L. & Gerstner, J. J. (2016). Using factor analysis in test construction. In K. Schweizer & C. DiStefano (Eds.), Principles and methods of test construction: Standards and recent advancements (pp. 26-51). Gottingen, Germany: Hogrefe.

Beaujean, A. A. (2015). John Carroll's views on intelligence: Bi-factor vs. higher- order models. *Journal of Intelligence, 3*, 121-136. doi:10.3390/jintelligence3040121

Beaujean, A. A., & Benson, N. F. (accepted). Theoretically-consistent cognitive ability test development and score interpretation. *Contemporary School Psychology*.

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods, 16*, 265–284

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., & Franić, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications.* (pp. 135--170). Charlotte, NC: Information Age Publishing.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061-1071. doi:10.1037/0033-295X.111.4.1061

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality, 80,* 796–846. doi.10.1111/j.1467-6494.2011.00749.x

Canivez, G. L. (2017). Review of the Woodcock-Johnson IV. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.), *The twentieth mental measurements yearbook* (pp. 875-882). Lincoln, NE: Buros Center for Testing.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY, US: Cambridge University Press. doi:10.1017/CBO9780511571312

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1-22. doi:10.1037/h0046743

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

Cohen, J. (1959) The factorial structure of the WISC at ages 7–6, 10–6, and 13–6. *Journal of Consulting Psychology*, 1959, *23*, 285–299.

Cormier, D. C., Bulut, O., McGrew, K. S., & Frison, J. (2016). The role of Cattell-Horn-Carol (CHC) cognitive abilities in predicting writing achievement during the school-age years. *Psychology in the Schools, 53*, 787-803. doi: 10.1002/pits.21945

Cormier, D. C., Bulut, O., McGrew, K. S., & Singh, D. (2017). Exploring the relations between Cattell-Horn-Carroll (CHC) cognitive abilities and mathematics achievement. *Applied Cognitive Psychology, 31*, 530-538. doi: 10.1002/acp.3350

Cormier, D. C., McGrew, K. S., Bulut, O., & Funamoto, A. (2017) Revisiting the relations between the WJ-IV measures of Cattell-Horn-Carroll (CHC) cognitive abilities and reading achievement during the school-age years. *Journal of Psychoeducational Assessment, 35*, 731-754. doi: 10.1177/0734282916659208

Deno, S.L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219-232.

Dombrowski, S. C. (2015). Exploratory bifactor analysis of the WJ III Achievement at school age via the Schmid-Leiman procedure. *Canadian Journal of School Psychology, 30*, 34-50. doi: 10.1177/0829573514560529

Dombrowski, S. C., McGill, R. J., & Canivez, G. L. (2018b). An alternative conceptualization of the theoretical structure of the Woodcock-Johnson IV Tests of Cognitive Abilities at school age: A confirmatory factor analytic investigation. *Archives of Scientific Psychology, 6*, 1-13. doi: 10.1037/arc0000039

Dombrowski, S. C., McGill, R. J., & Canivez, G. L. (2018a). Hierarchical exploratory factor analyses of the Woodcock-Johnson IV full test battery: Implications for CHC application in school psychology. *School Psychology Quarterly, 33*, 235-250. doi: 10.1037/spq0000221

Dombrowski, S. C., McGill, R. J., & Canivez, G. L. (2017). Exploratory and hierarchical factor analysis of the WJ IV Cognitive at school age. Psychological Assessment, 29, 294-407. doi: 10.1037/pas0000350

Dombrowski, S. C., DiStefano, C., & Noonan, K. (2004). Review of the Stanford-Binet, Fifth Edition. *Communiqué, 33*, 12-15.

Dombrowski, S. C., & Noonan, K. (2004). Review of the WISC-IV. *Communiqué, 33*, 35-38.

Fayers, P. M., & Hand, D. J. (2002). Causal variables, indicator variables and measurement scales: An example from quality of life. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 165*, 233-261.

Gaffney, T. W., Cudeck, R., Ferrer, E., & Widaman, K. F. (2010). On the factor structure of standardized educational achievement tests. *Journal of Applied Measurement, 11*, 384-

408.

Groth-Marnat, G., & Wright, A. J. (2016). *Handbook of psychological assessment* (6th ed.). Hoboken, NJ: Wiley.

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

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179 –185. doi: 10.1007/BF02289447

Jackson, J. S. H., & Maraun, M. (1996). The conceptual validity of empirical scale construction: The case of the sensation seeking scale. *Personality and Individual Differences, 21*, 103-110. doi:10.1016/0191-8869(95)00217-0

Jennrich, R. I., & Bentler, P. (2011). Exploratory bifactor analysis. *Psychometrika, 76*, 537-549. doi:10.1007/s11336-011-9218-4

Jennrich, R. I., & Bentler, P. M. (2012). Exploratory bifactor analysis: The oblique case. *Psychometrika, 77*, 442-454. doi:10.1007/s11336-012-9269-1

Jensen, A.R., & Weng, L. (1993). What is a good g? Intelligence 18, 231-258

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika,* 39, 31-36.

Kaufman, A. S. (1975). Factor analysis of the WISC-R at 11 age levels between 61/2 and 161/2 years. *Journal of Consulting and Clinical Psychology, 43*, 135-147. doi:10.1037/h0076502

Kaufman, A. S., & Kaufman, N. L. (2014). *Kaufman Test of Educational Achievement, Third*

Edition. Bloomington, MN: NCS Pearson.

Kaufman, A. S., & Lichtenberger, E. O. (2006). *Assessing adolescent and adult intelligence* (3rd ed.). Hoboken, NJ: Wiley.

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. doi:10.1016/j.intell.2012.01.009

Krause, M. S. (2005). How the psychotherapy research community must work toward measurement validity and why. *Journal of Clinical Psychology, 61*, 269-283. doi:10.1002/jclp.20020

Krause, M. S. (2012). Measurement validity is fundamentally a matter of definition, not correlation. *Review of General Psychology, 16*, 391-400. doi:10.1037/a0027701

Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (2006, April). *Scalability and the development of useful diagnostic scales*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Macmann, G. M., & Barnett, D. W. (1997). Myth of the master detective: Reliability of interpretations for Kaufman's "intelligent testing" approach to the WISC-III. *School Psychology Quarterly, 12,* 197-234. doi: 10.1037/h0088959

Maraun, M. D. (1998). Measurement as a normative practice: Implications of Wittgenstein's philosophy for measurement in psychology. *Theory & Psychology, 8*, 435-461. doi:10.1177/0959354398084001

Matarazzo, J. D., & Prifitera, A. (1989). Subtest scatter and premorbid intelligence: Lessons from the WAIS-R standardization sample. *Psychological Assessment, 1,* 186-191.

doi:10.1037/1040-3590.1.3.186

Mather, N., & Wendling, B. J. (2014). *Examiner's Manual. Woodcock-Johnson IV Tests of Achievement*. Rolling Meadows, IL: Riverside

McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment, 8,* 290-302. doi: 10.1177/073428299000800307

McDermott, P. A., Fantuzzo, J. W., Glutting. J. J., Watkins, M. W., & Baggaley, A. R. (1992). Illusions of meaning in the ipsative assessment of children's ability. *Journal of Special Education, 25,* 504-526. doi: 10.1177/002246699202500407

McGill, R. J., & Dombrowski, S. C. (2016). What does the WRAML2 core battery measure? Utilizing exploratory and confirmatory techniques to disclose higher order structure. *Assessment*. Advance online publication. doi: 10.1777/1073191116677799

McGill, R. J., & Dombrowski, S. C. (2017). School psychologists as consumers of research: What school psychologists need to know about factor analysis. *Communique, 46 (1)*.

McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). Technical Manual. Woodcock-Johnson IV. Rolling Meadows, IL: Riverside.

Muthén, L.K. and Muthén, B.O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén

Pearson. (2009). *Wechsler Individual Achievement Test* (3rd ed.). San Antonio, TX: Author.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47,* 667–696. doi:10.1080/00273171.2012.715555

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*, 544-559. doi:10.1080/00223891.2010.496477

Revelle, W. (2017). *psych: Procedures for psychological, psychometric, and personality research* (version 1.7.8) [computer software]. Evanston, IL: Northwestern University.

Reynolds, C. R. (1979). Factor structure of the Peabody Individual Achievement Test at five grade levels between grades one and 12. Journal of School Psychology, 17, 270-274.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: calculating and interpreting statistical indices. Psychological Methods, 21, 137-150. doi: 10.1037/met0000045

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53-61. doi:10.1007/BF02289209

Schneider, W. J., & 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* (3rd ed., pp. 99-144). New York: Guilford.

Schrank, F. A., Mather, N., & McGrew, K. S. (2014). *Woodcock-Johnson IV Tests of Achievement*. Rolling Meadows, IL: Riverside.

Shapiro, E. S. (2010). *Academic Skills Assessment* (2nd ed.). New York: Guilford Press.

Shinn, M.R. (Ed.). (1989) Curriculum-based measurement: Assessing Special Children, (pp 1-17). NY: Guilford Press.

Shojima, K., Otsu, T., Mayekawa, S.-i., Taguri, M., & Yanai, H. (2007). Factor structure of the National Center Test 2005 by the full-information pseudo-ML method. *Behaviormetrika, 34*, 131-156. doi:10.2333/bhmk.34.131

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 21,* 100-107.

Velicer, W. F. (1976). Determining the number of components form the matrix of partial correlations. *Psychometrika, 31*, 321-327. doi: 10.1007/BF02293557

Watkins, M. W. (2000). Cognitive profile analysis: A shared professional myth. *School Psychology Quarterly, 15,* 465-479. doi: 10.1037/h0088802

Watkins, M. W. (2003). IQ subtest analysis: Clinical acumen or clinical illusion? *The Scientific Review of Mental Health Practice, 2,* 118-141.

Watkins, M. W. (2006). Orthogonal higher order structure of the Wechsler Intelligence Scale for Children-Fourth Edition. *Psychological Assessment, 18,* 123-125. doi: 10.1037/1040-3590.18.1.123

Watkins, M. W., & Kush, J. C. (1994). Wechsler subtest analysis: The right way, the wrong way, or no way? *School Psychology Review, 23,* 640-651. Retrieved from http://www.nasponline.org

Williams, T. O., & Eaves, R. C. (2001). Exploratory and confirmatory factor analyses of the Woodcock reading mastery tests–revised with special education students. *Psychology in the Schools, 38*, 561-567.

Williams, T. O., Fall, A., Eaves, R. C., Darch, C., & Woods-Groves, S. (2007). Factor analysis of the KeyMath—Revised Normative Update Form A. *Assessment for Effective Intervention, 32*, 113-120. doi: 10.1177/15345084070320020201

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock Johnson III Tests of Achievement*. Rolling Meadows, IL: Riverside.

Table 1

		Grw						Mathematics				Writing				Cross-Domain Clusters					
		Reading	Broad Reading	Basic Reading Skills	Reading Comprehension	Reading Fluency	Reading Rate	Mathematics	Broad Mathematics	Math Calculation Skills	Math Problem Solving	Written Language	Broad Written Language	Basic Writing Skills	Written Expression	Academic Skills	Academic Fluency	Academic Applications	Academic Knowledge	Phoneme-Grapheme Knowledge	Brief (or Broad) Achievement
Standard Battery	**Subtest/CHC Factor**																				
	Letter-Word Identification (Grw)	✓	✓	✓												✓					✓
	Applied Problems (Gq/Gf)							✓	✓		✓							✓			✓
	Spelling (Grw)											✓	✓	✓		✓					✓
	Passage Comprehension (Grw)	✓	✓		✓													✓			◊
	Calculation (Gq)							✓	✓	✓						✓					◊
	Writing Samples (Grw)											✓	✓		✓			✓			◊
	Word Attack (Grw/Ga)			✓																✓	
	Oral Reading (Grw)					✓															
	Sentence Reading Fluency (Grw/Gs)		✓			✓	✓										✓				◊
	Math Facts Fluency (Gq/Gs)								✓	✓							✓				◊
	Sentence Writing Fluency (Grw/Gs)												✓		✓		✓				◊
	Reading Recall (Grw/Glr)				✓																

Extended Battery	Number Matrices (Gf)										✓										
	Editing (Grw)													✓							
	Word Reading Fluency (Grw/Gs)						✓														
	Spelling of Sounds (Grw/Ga)																			✓	
	Reading Vocabulary (Grw/Gc)				•																
	Science (Gc)																		✓		
	Social Studies (Gc)																		✓		
	Humanities (Gc)																		✓		

✓ Tests required to create the cluster listed.

• Additional test required to create an extended version of the cluster listed

◊ Additional tests required to create the Broad Achievement cluster.

Note: Grw = Reading/Writing, Gq = Quantitative Reasoning, Gc = Academic Knowledge/Crystallized Ability, Gs = Academic Fluency/Processing Speed. Ga=Auditory Processing.

From W*oodcock-Johnson IV™ (WJ IV™)* Technical Manual. Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved. Used by permission of the publisher. Any further duplication is strictly prohibited unless written permission is obtained from Houghton Mifflin Harcourt Publishing Company.

Table 2

Factor Extraction Results

Extraction Methods	Age 9 to 13	Age 14 to 19
Empirical BIC	6	5
MAP	3	4
PA, Factors (50th and 95th%ile)	6 & 6	5 & 6
PA, Components (50th and 95th%ile)	3	3
Scree	4–5	4–5
Kaiser (Eigenvalue >1)	3	3

Note: BIC=Bayesian Information Criterion; MAP=Minimum average partial test; PA=Parallel Analysis

Table 3

Exploratory Bifactor Analysis Variance Apportionment of the Woodcock-Johnson Tests of Achievement-Fourth Edition (Ages 9-13) Six Factor Solution (one general & 5 group factors)

	Factor									
Test	ACH-*g*	Ga	Gq	Grw	Gs	Gc	h^2	u^2	e^2	s^2
Letter Word Identification (Grw)	.87	.22	-.07	.08	-.02	-.02	.81	.19	.06	.13
Applied Problems (Gq/Gf)	.70	-.03	**.35**	.02	.00	.23	.67	.34	.08	.26
Spelling (Grw)	.62	**.38**	.04	-.07	.02	-.02	.54	.46	.12	.34
Passage Comprehension (Grw)	.80	-.01	-.10	**.27**	-.07	.00	.72	.28	.11	.17
Calculation (Gq)	.74	.01	**.51**	.18	.02	.08	.84	.16	.07	.09
Writing Samples (Grw)	.60	.22	-.01	**.44**	.02	.02	.60	.40	.10	.30
Word Attack (Grw/Ga)	.69	**.48**	-.17	.06	-.09	.02	.75	.25	.10	.15
Oral Reading (Grw)	.78	.03	-.16	-.11	.05	-.14	.67	.33	.04	.29
Sentence Reading Fluency (GrwGs)	.75	.01	.00	.04	**.50**	-.08	.82	.19	.06	.13
Math Fact Fluency (Gq/Gs)	.66	.02	**.37**	-.05	**.39**	-.09	.74	.26	.04	.22
Sentence Writing Fluency	.71	-.01	.08	.12	**.40**	-.09	.68	.32	.20	.12
Reading Recall (Grw/Glr)	.64	-.03	.03	**.33**	-.01	-.09	.53	.47	.08	.39
Number Matrices (Gf)	.57	-.07	**.31**	-.14	.02	.18	.48	.52	.08	.44
Editing (Grw)	.83	-.03	-.05	-.08	.02	.06	.71	.29	.09	.20
Word Reading Fluency (Grw/Gs)	.63	-.06	.01	-.01	**.56**	-.02	.72	.28	.08	.20
Spelling of Sounds (Grw/Ga)	.62	**.38**	.04	-.07	.02	-.02	.54	.46	.12	.34
Reading Vocabulary (Grw/Gc)	.84	-.21	-.16	-.03	-.03	.21	.82	.18	.12	.06
Science (Gc)	.54	.06	.01	.10	-.07	**.70**	.79	.21	.16	.05
Social Studies (Gc)	.60	.03	.02	-.04	-.01	**.53**	.64	.36	.13	.23
Humanities (Gc)	.56	-.09	.01	-.20	-.06	**.49**	.61	.39	.13	.26
Common Variance (%)	.72	.04	.05	.04	.07	.09	.70	.30	.10	.21
Total Variance (%)	.50	.03	.04	.03	.05	.06				

$\omega_{H}/\ \omega_{HS}$	.90	.13	.21	.17	.28	.33
H	.96	.36	.37	.31	.54	.63

Note. ACH-*g* = general achievement, Grw = Reading/Writing, Gq = Quantitative Reasoning, Gc = Academic Knowledge/Crystallized Ability, Gs = Academic Fluency/Processing Speed. Ga=Auditory Processing. h^2 = communality; u^2 = uniqueness; e^2 =error (1-reliability); Reliability estimates from McGrew, LaForte, & Schrank (2014); s^2 = subtest specific variance (u^2-error); ω_H = omega hierarchical; ω_{HS_h} = omega-hierarchical subscale. H= Index of construct replicability.

Table 4

Exploratory Bifactor Analysis Variance Apportionment of the Woodcock-Johnson Tests of Achievement-Fourth Edition (Ages 14-19) Six Factor Solution (one general & 5 group factors)

	Factor									
Test	ACH-*g*	Gq	Grw	Ga	Gs	Gc	h^2	u^2	e^2	s^2
Letter Word Identification (Grw)	.88	-.11	-.02	.18	-.01	-.06	.82	.18	.06	.12
Applied Problems (Gq/Gf)	.75	**.38**	.02	-.06	.01	.24	.77	.24	.08	.16
Spelling (Grw)	.86	-.01	-.19	.12	.10	-.01	.80	.20	.08	.12
Passage Comprehension (Grw)	.83	-.07	**.28**	.01	-.05	-.01	.77	.23	.11	.12
Calculation (Gq)	.76	**.53**	.12	.02	.01	.10	.89	.11	.07	.04
Writing Samples (Grw)	.65	.00	**.36**	.24	-.04	.00	.60	.40	.10	.30
Word Attack (Grw/Ga)	.72	-.13	.02	**.45**	-.16	.01	.76	.24	.10	.14
Oral Reading (Grw)	.78	-.12	-.05	.02	.07	-.12	.64	.36	.04	.32
Sentence Reading Fluency (GrwGs)	.72	-.04	.07	-.01	**.53**	-.04	.81	.19	.06	.13
Math Fact Fluency (Gq/Gs)	.61	**.41**	-.06	.02	**.42**	-.04	.72	.28	.04	.24
Sentence Writing Fluency	.68	.06	.08	.00	**.40**	-.09	.64	.36	.20	.16
Reading Recall (Grw/Glr)	.59	.02	**.32**	-.03	.05	-.17	.49	.51	.08	.43
Number Matrices (Gf)	.61	**.30**	-.14	-.06	.01	.21	.53	.47	.08	.39
Editing (Grw)	.83	.00	-.12	-.07	.02	.04	.72	.28	.09	.19
Word Reading Fluency (Grw/Gs)	.58	.03	-.03	-.06	**.59**	-.01	.69	.31	.08	.23
Spelling of Sounds (Grw/Ga)	.67	.06	.00	**.33**	.03	-.03	.57	.43	.12	.31
Reading Vocabulary (Grw/Gc)	.86	-.12	.00	-.21	-.02	.20	.84	.16	.12	.04
Science (Gc)	.65	.02	.08	.03	-.11	**.56**	.76	.24	.16	.08
Social Studies (Gc)	.66	.05	-.04	.06	-.03	**.51**	.71	.29	.13	.16
Humanities (Gc)	.63	.00	-.16	-.07	-.04	**.51**	.68	.32	.13	.19
Common Variance (%)	.74	.05	.03	.03	.07	.07	.71	.29	.10	.19
Total Variance (%)	.52	.04	.02	.02	.05	.05				
$\omega_{H/}\ \omega_{HS}$	.93	.22	.14	.13	.32	.26				

H	.96	.40	.26	.32	.57	.54

Note. ACH-*g* = general achievement, Grw = Reading/Writing, Gq = Quantitative Reasoning, Gc = Academic Knowledge/Crystallized Ability, Gs = Academic Fluency/Processing Speed. Ga=Auditory Processing. h^2 = communality; u^2 = uniqueness e^2=error (1-reliability); Reliability estimates from McGrew, LaForte, & Schrank (2014); s^2 = subtest specific variance (u^2-error); ω_H = omega hierarchical; ω_{HS} = omega-hierarchical subscale. H= Index of construct replicability.

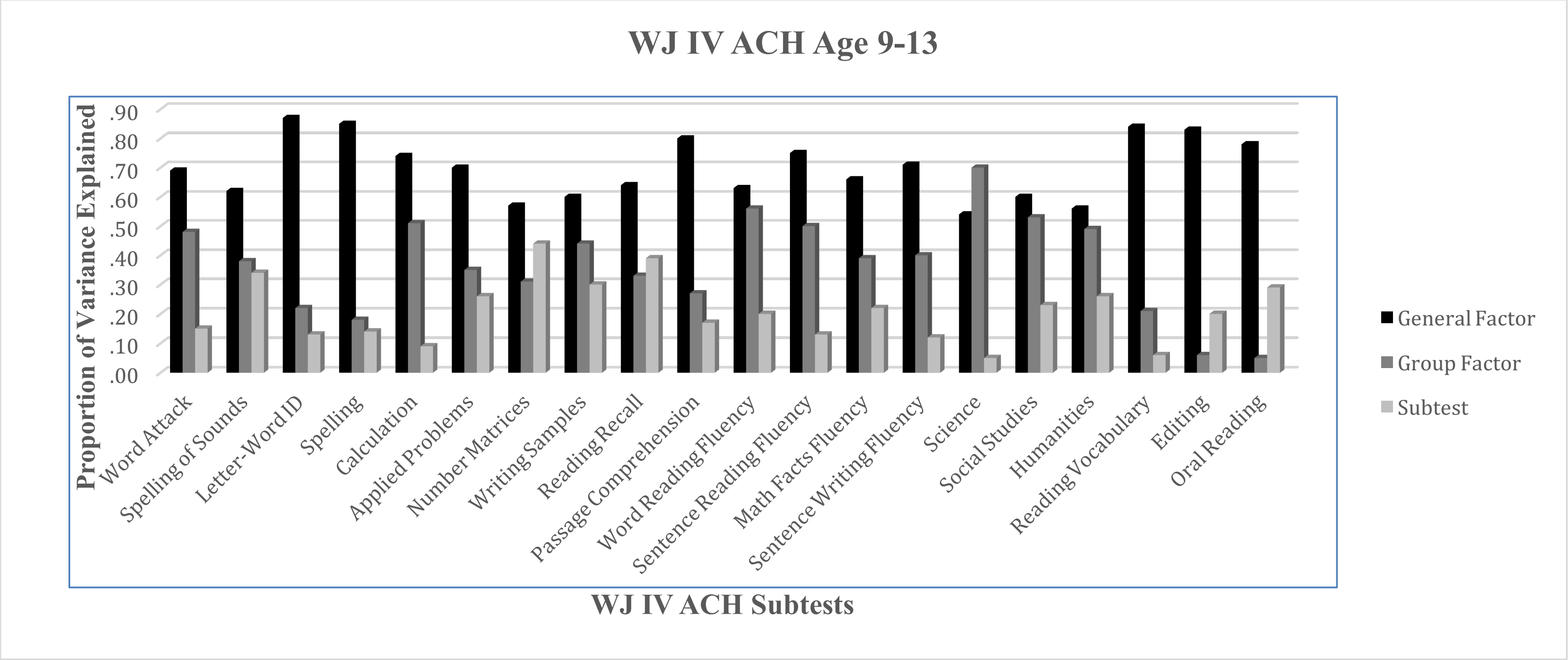

Figure 1. Variance apportionment among general, group and specific WJ IV ACH age 9-12 subtests.

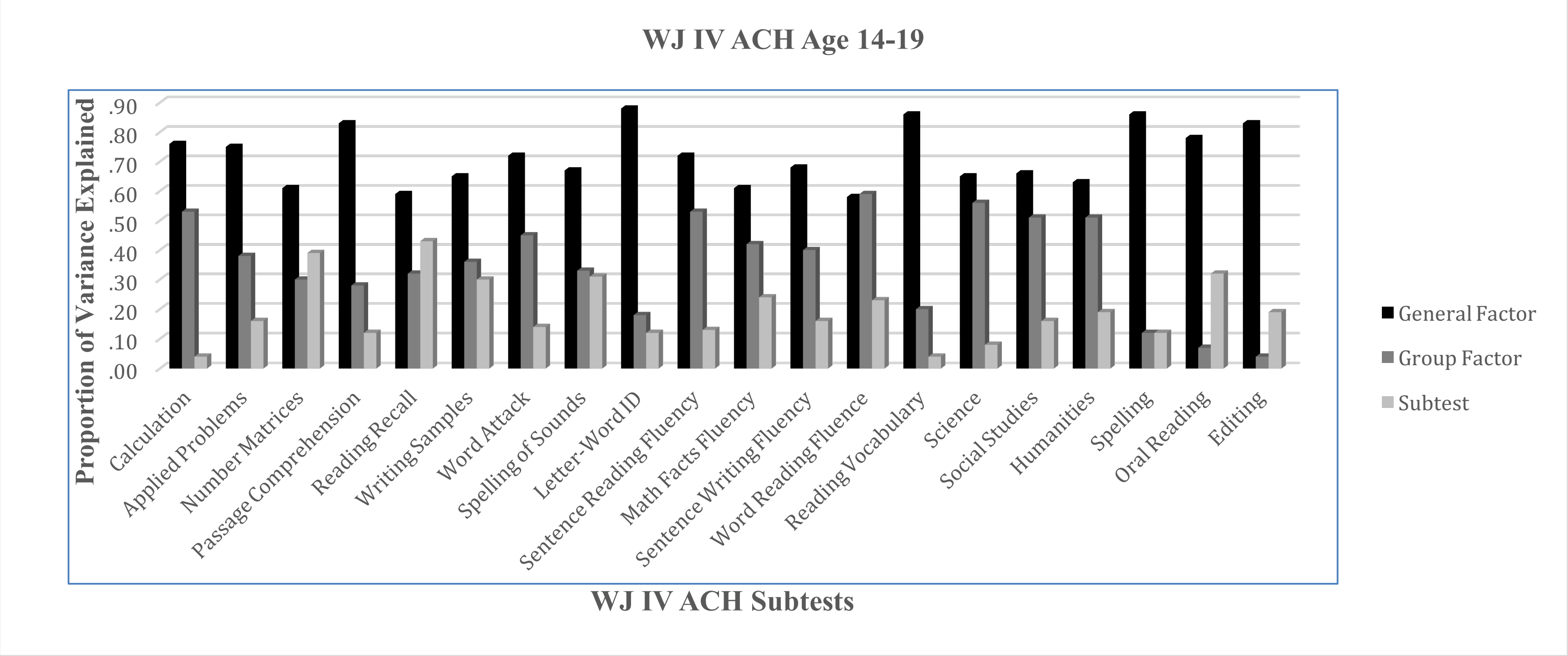

Figure 2. Variance apportionment among general, group and specific WJ IV ACH age 14-19 subtests.