




Challenges to the Cattell-Horn-Carroll Theory: Empirical, Clinical, and Policy Implications

Gary L. Canivez ^a and Eric A. Youngstrom^b



^aPsychology, Eastern Illinois University; ^bPsychology and Neuroscience, University of North Carolina at Chapel Hill

ABSTRACT

The Cattell-Horn-Carroll (CHC) taxonomy of cognitive abilities married John Horn and Raymond Cattell's Extended Gf-Gc theory with John Carroll's Three-Stratum Theory. While there are some similarities in arrangements or classifications of tasks (observed variables) within similar broad or narrow dimensions, other salient theoretical features and statistical methods used for examining and supporting them are in direct opposition. In this article, the theoretical disagreements between Carroll and Cattell-Horn and theoretical incongruities between their models are delineated, which raises substantive challenges to CHC. Additionally, there are practical and substantial measurement obstacles that further threaten *practical* application of CHC. We conclude that the problems are due to some fundamental differences that likely will not change, so call for an annulment of this arranged but unhappy marriage.

Authors and publishers of most major contemporary tests of intelligence claim alignment with the Cattell-Horn-Carroll (CHC) model of cognitive abilities in test development, revision, and in score interpretations (c.f., Elliott, 2007a; 2007b; Kaufman & Kaufman, 2004; McGrew & Woodcock, 2001; McGrew, LaForte, & Schrank, 2014; Roid, 2003a; 2003b; Schrank, McGrew, & Mather, 2014a; 2014b; Wechsler, 2014a; 2014b).¹ The CHC model quickly became popular without much critical appraisal, although some pointed out some inherent problems with it – particularly its application in cross-battery assessment (Glutting, Watkins, & Youngstrom, 2003). CHC was first referred to as a *taxonomy* for classifying subtests or tasks into broad categories or factors, and often narrow abilities assignments were based on “logical or expert-consensus” (McGrew, 2014). However, now it is commonly referred to as a theory, not a taxonomy (McGrew & Woodcock, 2001; Schneider & McGrew, 2012); although Deary (2001) would disagree and refer to it as a taxonomy. Many inferences about subtest and composite scores from contemporary intelligence tests might appear to be in line with a CHC model, but there are considerable psychometric problems involved in interpreting scores in a CHC framework that are not adequately (or at all) addressed in test technical or interpretive manuals (e.g., Elliott, 2007b; McGrew et al., 2014; McGrew & Woodcock, 2001; Wechsler, 2014b).

Most modern intelligence tests follow Wechsler's design of his Bellevue test. They require samples of a wide variety of behaviors and produce multiple scores to interpret, with the overarching goal to aid in clinical decision-making instead of the measurement of psychological attributes. While many tests contain a score representing average performance across the tasks, the suggested interpretive practices *deemphasize* the role of any score representing *g* (Spearman, 1927, 1931) in favor of broad

CONTACT Gary L. Canivez  gcanivez@eiu.edu  Psychology, Eastern Illinois University, 600 Lincoln Avenue, Charleston, IL 61920, USA

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/home.

¹We use the term *intelligence* throughout this article to refer to the class of attributes within the general domain of cognitive ability, not any particular attribute within that domain.

or narrow ability scores as well as ipsative comparisons or profile examinations (i.e., strengths/weaknesses).

For example, McGrew and Flanagan (1998) wrote that score interpretations should focus on “unique *Gf-Gc pattern of abilities* of individuals that in turn can be related to important occupational and achievement outcomes and other human traits,” and purposefully eschewed anything related to *g* because it “has little practical relevance to cross-battery assessment and interpretation” (p. 14). Moreover, “a global composite intelligence test score is at odds with the underlying *Gf-Gc* cross-battery philosophy” (p. 382) and the focus on cognitive profiles “uncovers the individual skills and abilities that are more diagnostic of learning and problem-solving processes than a global IQ score” (p. 383). The McGrew and Flanagan appeal to “prominent psychological assessment spokespersons,” Kaufman and Lezak, and “beliefs” about the utility of “lower levels (viz., stratum I and II *Gf-Gc* abilities)” (p. 382) also denied *g* from the very beginning.

CHC-based interpretive approaches emphasize broad abilities over a global score (i.e., WJ IV GIA, WISC-V FSIQ). Appeal to theoretical association of interpretive practices is inadequate. Such association with CHC is not, *ipso facto, evidence* for the scores’ validity or, more importantly, diagnostic and treatment utility. McGill and Dombrowski (2019) described CHC theory as a combination or marriage of Horn and Cattell’s Extended *Gf-Gc* theory (E *Gf-Gc*; Horn & Blankson, 2005; Horn & Noll, 1997) and Carroll’s Three-Stratum Theory (3S; Carroll, 1993). There is similarity in arrangement or classification of subtest tasks within similar broad abilities, but that appears to be the main – and perhaps only – line of congruence; other salient theoretical features are in direct opposition, and there is discordance in the statistical methods that have been used to evaluate the models.

Horn and Blankson (2005) famously wrote “the extended theory of fluid and crystallized (*Gf* and *Gc*) cognitive abilities is wrong” because “all scientific theory is wrong” (p. 41). While this is undoubtedly true, some theories may be more wrong than others. Although there are theoretical aspects of intelligence that are interesting and important, there are also *practical* aspects of intellectual assessment with our present tests that may not adequately capture the underlying attributes. It is one thing to assert that a test follows a model, but it something entirely different to examine if data conform to strong predictions from the model (Roberts & Pashler, 2000). Also crucial are examinations of the reliability, validity, and diagnostic and treatment utility. To that end, in this article, we examine the evidence included in major test technical manuals claiming CHC support as well as independent analyses of standardization sample data challenging the overly optimistic conclusions offered by the test publishers for what their tests are measuring. In addition, we also include a review of relevant research challenging the reliability, validity, and utility of CHC-based interpretive practices.

1. Theoretical Relations of Horn-Cattell and Carroll

Horn denied the existence of *g*, referring to it as an artifact of positive manifold and factor analysis. The Horn-Cattell Extended *Gf-Gc* theory (Horn & Blankson, 2005; Horn & Noll, 1997) included eight broad ability factors (*Gc, Gf, Gsm, Gln, Gs, Gv, Ga, Gq*), but *did not* include a dimension related to *g*, despite the substantial covariance among the eight broad factors (see Figure 1). Thus,

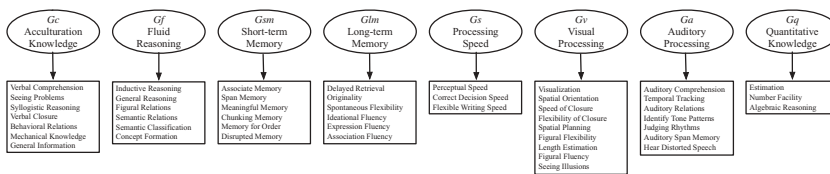


Figure 1. Horn-Cattell extended *Gf-Gc* theory.

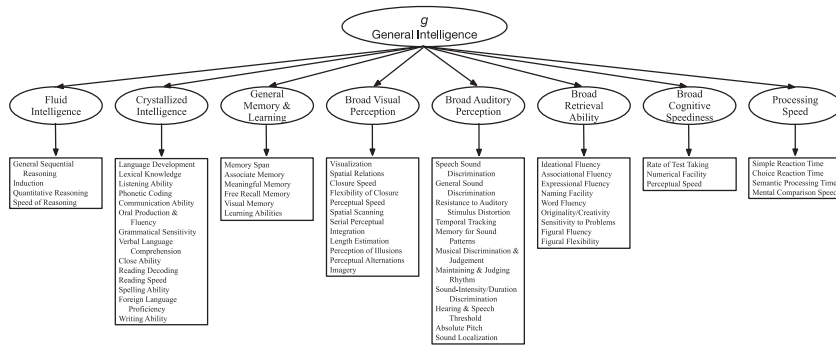


Figure 2. Higher-order representation of Carroll's three-stratum theory.

the Horn-Cattell theory is similar in structure and conceptualization to Thurstone's (1938) Primary Mental Abilities, although *Gf-Gc* attributes are broader than Thurstone's (Horn, 1991).

Carroll (1993), in his re-analyses of hundreds of data sets described a model that included broad ability factors (Stratum II) similar to Horn-Cattell, but also included the *g* dimension (Stratum III). While most representations of Carroll's theoretical structure present it as a higher-order structure with a general factor at the apex (Figure 2), this is not what Carroll intended. Instead, Carroll (1996) noted that his model was largely just an extension of Spearman's two-factor model, similar to a bifactor structure (Holzinger & Swineford, 1937) where *g* and the broad abilities have direct and simultaneous influences on cognitive tasks (see Figure 3) and are at the same level of inference (Canivez, 2016). So, while both Horn-Cattell and Carroll models have indicators associated with broad ability dimensions, only Carroll's includes *g*.

CHC, as originally conceived by McGrew and Woodcock (2001) and later extended (Schneider & McGrew, 2012), combined the Horn-Cattell *E Gf-Gc* model with Carroll's 3S model in order to classify and align subtest indicators within broad ability dimensions similar to both models. CHC, as well as interpretive guidelines for contemporary intelligence tests (c.f., Elliott, 2007b; McGrew et al., 2014; McGrew & Woodcock, 2001; Wechsler, 2014b), however, seem to only grudgingly acknowledge a general higher-order factor (and as such the influence of this general factor is fully mediated by the first-order broad abilities), but deemphasize *g* with most analyses, discussion, and expansive interpretations of the influences or effects of the broad CHC abilities (see also Cormier, Bulut, McGrew, & Frison, 2016; Cormier, Bulut, McGrew, & Singh, 2017; Cormier, McGrew, Bulut, & Funamoto, 2017).

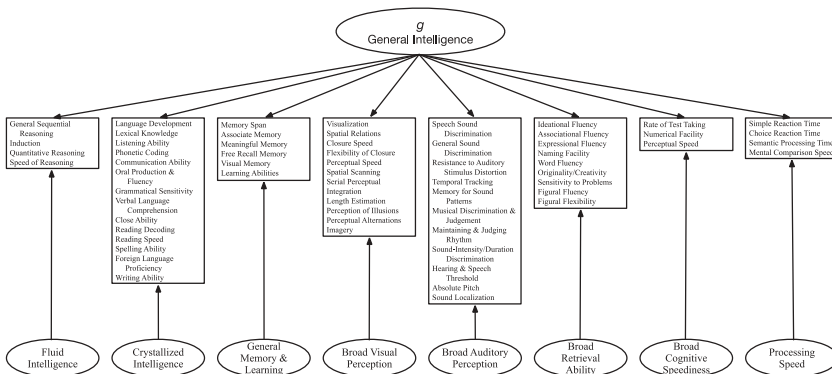


Figure 3. Bifactor representation of Carroll's three-stratum theory.

The Horn-Cattell approach (denial of g) differs substantively from that of Carroll (cognitive tests appear to be primarily, but not solely, influenced by g), which makes for strange bedfellows and a mighty uncomfortable arranged marriage. The Horn-Cattell Extended Gf-Gc theory appears unnecessary given Carroll's Three-Stratum Theory (see Cucina & Howardson, 2017), and Carroll (2003, p. 18) himself noted in his final publication that McGrew and Woodcock (2001) introduced "a so-called CHC (Cattell-Horn-Carroll) theory of cognitive abilities that supplemented Horn's Gf-Gc theory with essentially a three-stratum theory..." and that "Even though I was to some extent involved in this change (as an occasional consultant to the authors and publishers), I am still not quite sure what caused or motivated it." In regard to g , he wrote:

Horn's comment suggests that he conveniently forgets a fundamental principle on which factor analysis is based (a principle of which he is undoubtedly aware) – that the nature of a single factor discovered to account for a table of inter-correlations does not necessarily relate to special characteristics of the variables involved in the correlation matrix; relates only to characteristics or underlying measurements (latent variables) that are common to those variables. I cannot regard Horn's comment as a sound basis for denying the existence of a factor g , yet he succeeded in persuading himself and many others to do exactly this for an extended period of years. (p. 19)

Carroll and Horn do not appear to be willing participants in this arranged marriage given their apparent disagreements and they themselves do not provide written statements or agreement regarding such union for a CHC. Cattell too, although passing away before development of CHC, would likely have been opposed to any model that included g (e.g., Cattell, 1987). It is particularly noteworthy that in a podcast describing CHC theory and forthcoming CHC changes, McGrew (2018a; see also McGrew, 2018b) stated "...there is no g in these models because... I think g is a statistical artifact, but that's something for another time" (31:04–31:12). This is an identical position to that of Horn and antithetical with Carroll, so it is perplexing why Carroll continues to be included in "so-called CHC" given his position regarding g . McGill and Dombrowski (2019) noted that before passing away Carroll was planning on further distancing himself from CHC because of his disagreement about the primacy of g (McGrew, 2005, p. 174 [11]). Based on Carroll's published comments on CHC and the substantive theoretical differences between Carroll and Horn-Cattell it appears that it is time to grant an *annulment* to the arranged marriage of their rival theories.

2. Problems with Tests Purporting to Assess Broad Attributes

In addition to the fundamental incongruities between the Horn-Cattell E Gf-Gc and Carroll's 3S, there are a number of additional psychometric problems and concerns regarding the intelligence tests supposedly constructed to assess the amalgam of broad attributes in CHC. These problems have substantial *practical* implications and limitations for interpretation and comparison of scores purportedly aligned to broad CHC attributes. These problems are enumerated and expanded upon to articulate *why* interpretations based primarily on broad attributes promoted in major intelligence tests are extremely problematic (regardless of CHC linkages), many of which lack basic psychometric fitness for confident interpretation. For illustrative purposes, examples from the Woodcock-Johnson III Tests of Cognitive Abilities (WJ III Cognitive; Woodcock, McGrew, & Mather, 2001), Woodcock-Johnson Tests of Cognitive Abilities, Fourth Edition (WJ IV Cognitive; Schrank, McGrew, & Mather, 2014b) and the Wechsler Intelligence Scale for Children-Fifth Edition (WISC-V; Wechsler, 2014a) are presented because the WJ III and WJ IV were designed to be *the* reference instruments reflecting CHC, while the WISC-V (one of the most frequently administered intelligence tests) was designed with specific intention to measure five of the broad CHC attributes.

2.1. Test Structure Replication and Broad Factor Identification

2.1.1. WJ III

The WJ III Cognitive (Woodcock et al., 2001) was the first instrument constructed to represent the CHC model. Assessment of the factor structure of the WJ III Cognitive reported in the manual (McGrew & Woodcock, 2001) relied exclusively on confirmatory factor analyses (CFA) and claimed alignment with CHC. Later, Dombrowski conducted a series of independent examinations into the factor structure of the instrument (Dombrowski, 2013, 2014a, 2014b, 2015; Dombrowski & Watkins, 2013) using the exploratory factor analytic (EFA) methods and was unable to replicate the structure reported in the WJ III manual. EFA is the preferred approach when there is uncertainty about the number of factors underlying the indicator variables. If a CFA starts off with the wrong number of factors, not only can it bias the loadings, but the model fit indices do not provide a clear algorithm that will recover the correct model.

Dombrowski and Watkins (2013) analyses showed that the subtests and WJ structure *did not* fully align with CHC theory. For the full test battery (i.e., including the Cognitive and Achievement subtests), they were only able to capture six (ages 9–13) or five factors (ages 14–19) – *not* the nine purported by CHC. Dombrowski's (2013) examination of the WJ III Cognitive using EFA found that the WJ III appeared to measure only four factors for ages 9–13 (combined Gf/Gv, Gc, Gs, and Glr) and only three factors for ages 14–19, (Gc, Gs, Gsm) – *not* the seven specified by CHC. Similar results were later obtained by Strickland, Watkins, and Caterino (2015). Thus, there were serious questions about the adequacy of the WJ III to assess the hypothesized CHC attributes it was designed to measure. Since the WJ III was custom built to represent CHC, failure to provide strong support for all the CHC attributes also raises questions about the adequacy of CHC itself.

2.1.2. WJ IV

The WJ IV Cognitive (Schrack et al., 2014b) is the latest version of the cognitive assessment battery within the Woodcock-Johnson IV (WJ IV; Schrank et al., 2014a), and the WJ IV *Technical Manual* (McGrew et al., 2014) indicated that the WJ IV Cognitive was designed to measure a hierarchically ordered general intellectual ability (i.e., *g*) and seven lower-order broad ability factors of Comprehension-Knowledge (*Gc*), Fluid Reasoning (*Gf*), Short-Term Working Memory (*Gwm*), Cognitive Processing Speed (*Gs*), Auditory Processing (*Ga*), Long Term Retrieval (*Glr*), and Visual-Processing (*Gv*). Again, the organization combines Carroll's (1993) emphasis on *g* with Horn and Cattell (1966) emphasis on broad abilities, even though neither camp agreed to the union. As noted by Dombrowski, McGill, and Canivez (2017, 2018a), the factor structure of the WJ IV Cognitive was not examined separately from the full battery, inflating the number of indicators and also adding additional constructs. Canivez (2017) and Dombrowski et al. (2017; 2018a, 2018b) articulated numerous criticisms of psychometric evidence for the WJ IV structure provided in its *Technical Manual* (McGrew et al., 2014). The CFA fit statistics presented in the WJ IV *Technical Manual* were particularly poor and could not be judged as remotely adequate in support of the hypothesized CHC model presented for the full battery. For each age group the best fitting initial models and cross-validation models did not have model fit values that even approached the levels considered adequate (e.g., Hu & Bentler, 1999). Thus, though the proposed final WJ IV models may have been the best among the models tested, they clearly were not well-fitting models.

Due to inadequate procedures and results and the lack of separate analyses for the WJ IV Cognitive instrument, Dombrowski et al. (2017) applied EFA to determine how many viable factors might emerge from the Cognitive subtests. None of the factor extraction criteria suggested seven factors – the number of CHC attributes the test was constructed to assess. When seven factors were forced, multiple factors were insufficiently defined, having only one (or none!) salient loadings, rendering it unacceptable and failing to support CHC. The most plausible structural models for the 9–13 and 14–19 age groups included *g* along with four factors representing broad attributes. Had

McGrew et al. (2014) considered such EFA results, the final WJ IV Cognitive structure and scores provided to interpret could have been significantly different.

Dombrowski, McGill, and Canivez (2018a) conducted an independent follow-up study of the structural validity of the WJ IV Cognitive using CFA, filling a gap as the *Technical Manual* did not provide separate analyses for the cognitive subtests. In addition to testing the adequacy of the hypothetical seven CHC Cognitive factor model, Dombrowski et al. also examined Woodcock's cognitive processing model (Taub & McGrew, 2014) and the four-factor models suggested by Dombrowski et al. (2017). Additionally, oblique and rival bifactor representations of all models were tested, as McGrew et al. also did not consider bifactor structures. In both the 9–13 and 14–19 age groups, the hypothesized seven-factor higher-order CHC-based model promoted in the technical manual was unacceptable since it produced Heywood cases (i.e., produced negative residual variance estimates). The oblique seven-factor models (no *g*) also failed to provide viable results. As with EFA, CFA results appeared to favor four factors representing broad abilities along with a factor representing *g* (i.e., bifactor structure). Thus, while the WJ IV provides scoring for seven cognitive factors, independent assessment does not support such configuration. As with the WJ III, since the WJ IV was custom built to represent CHC, these results also undermine the claimed support for the CHC model itself.

2.1.3. WISC-V

The Wechsler scales have historically been atheoretical, although perhaps a better description is that they have been designed to comport with multiple (even competing) models of intelligence (c.f., Littell, 1960). As with its predecessors, the WISC-V was designed to align with a variety of intelligence models, including CHC. Thus, the authors divided what was formerly called the Perceptual Reasoning index into separate Visual-Spatial (VS [*Gv*]) and Fluid Reasoning (FR [*Gf*]) scores (Wechsler, 2014b). The claim by the publisher to assess five broad abilities and the “support” provided in the WISC-V *Technical and Interpretive Manual* (Wechsler, 2014b) was based on questionable statistical and methodological practices (see Beaujean, 2016; Canivez & Watkins, 2016). The lack of details about procedures and important results from the factor models led to independent analyses. Canivez, Watkins, and Dombrowski (2016) subjected the 16 WISC-V subtests' correlations for the total standardization sample to an EFA.² Results showed that no extraction criteria recommended extraction of five factors. Forced extraction of five factors produced an inadequate fifth factor with only one salient subtest pattern coefficient, while simultaneously dividing covariance in Matrix Reasoning such that it did not saliently load on any factor. The four-factor extraction produced acceptable structure in subtest alignment that was identical to the WISC-IV (Watkins, 2006): a dominant general factor and four smaller group factors. The attempt to separate Block Design (BD) and Visual Puzzles (VP) into a VS (*Gv*) factor and Matrix Reasoning (MR), Figure Weights (FW), and Picture Concepts (PCon) into a FR (*Gf*) factor was unsuccessful; BD, VP, MR, and FW formed a factor resembling the former Perceptual Reasoning factor. Canivez, Watkins, and Dombrowski (2017) followed up this study with independent CFA and found *all* the publisher specified models with five group factors produced serious problems with the results (e.g., Heywood cases). A bifactor model with *g* and four group factors was judged best, verifying the results from the EFA and, again, suggesting that alignment to CHC (viz., separation of VS [*Gv*] and FR [*Gf*]) was not supported. Subsequent independent analyses of WISC-V structure with smaller standardization sample age groups also failed to support five factors (Canivez, Dombrowski, & Watkins, 2018; Dombrowski, Canivez, & Watkins, 2018).

Results from the WJ III, WJ IV, and WISC-V all converged on a common result: despite claims by publishers and authors that the scores represent CHC broad abilities, independent replication of cognitive test structures has been problematic. The subtests do not always load on their intended factors and sometimes fail to load on any factor. Further, there are consistently fewer *viable* factors

²NCS Pearson declined access to the standardization data for independent analyses.

identified by independent researchers than what publishers promote and for which they provide scores (Canivez et al., 2018; Canivez & Watkins, 2016; Canivez et al., 2016; 2017; Dombrowski, 2013, 2014a, 2014b; Dombrowski, 2015; Dombrowski et al., 2018, 2017, 2018a, 2018b; Dombrowski & Watkins, 2013). Thus, clinicians are very likely providing interpretations of scores and making inferences about individuals' performances that are not empirically supported and may well lead to erroneous clinical decisions. Interpreting extra scores increase the chances of false positive (Type I) errors (Silverstein, 1993). Interpreting scores with inadequate specification or reliable variance at best attenuates the power to detect differences (increasing the Type II error rate); perhaps more problematic, however, is that the numerals provided for scores related to these factors do not actually represent measurement of any attribute – they are only indicative of some statistical chimera that is irrevocably tied to a specific intelligence test.

2.2. Unique Broad Factor Contribution

Intelligence tests often provide an omnibus, full-scale score (i.e., WJ IV GIA, WISC-V FSIQ) and scores representing broad abilities that are composites of subtests that ostensibly represent a given broad ability. The authors/publishers then represent these scores as being equally interpretable. Practitioners are frequently unaware that scores representing broad abilities conflate sources of variance, which substantially complicates their interpretation. Thus, variance in these scores is partly due to *g* and partly due to the broad ability. While the proportion of variance due to each source can be decomposed for a sample of respondents, it cannot be known for a given individual how much of the obtained score comes from the different attributes (Oh, Glutting, Watkins, Youngstrom, & McDermott, 2004).

For most intelligence tests, the broad ability scores are not *pure* measures of the broad ability and should not be interpreted to reflect *only* that broad ability. In fact, most such scores contain substantially more *g* variance than broad ability variance! Carroll's (1993) use of second-order EFA followed by SL transformation was done to better understand the relative contributions of *g* and the group factors in measurement of cognitive abilities. The SL procedure "...preserves the desired interpretation characteristics of the oblique solution, but also discloses the hierarchical structuring of the variables" (Schmid & Leiman, 1957, p. 53). This is why Carroll (1995) insisted on SL orthogonalization of higher-order models:

I argue, as many have done, that from the standpoint of analysis and ready interpretation, results should be shown on the basis of orthogonal factors, rather than oblique, correlated factors. I insist, however, that the orthogonal factors should be those produced by the Schmid-Leiman (1957) orthogonalization procedure, and thus include second-stratum and possibly third-stratum factors. (p. 437)

The benefit of SL in EFA (and bifactor models in EFA and CFA) is understanding the relative contributions of the factors, but an added benefit is in using such orthogonalized factors in estimating model-based reliability estimates.

2.3. Model-Based Reliabilities

Internal consistency reliability estimates provided in intelligence test manuals are likely inappropriate based on violations of assumptions for their estimation (Raykov, 1997). Omega (ω) coefficients (McDonald, 1999) have more realistic assumptions, so are more appropriate (Watkins, 2017). Coefficients omega-hierarchical (ω_H) and omega-hierarchical subscale (ω_{HS}) are estimates of reliability for composite scores produced by the particular subtest groupings (Reise, 2012; Rodriguez, Reise, & Haviland, 2016; Watkins, 2017). The ω_H coefficient is reliability for a general score after removing variability from the group factors. The ω_{HS} coefficient is the reliability estimate for a specific broad ability score after removing variability from the general and all other group factors (Brunner, Nagy, & Wilhelm, 2012; Reise, 2012). ω_H and ω_{HS} can be estimated from EFA SL results,

bifactor model estimates, or decomposed variance estimates from CFA higher-order models. ω_H and ω_{HS} coefficients should exceed .50, but .75 might be preferred (Reise, 2012; Reise, Bonifay, & Haviland, 2013) for confident scale interpretation.

Factor analytic studies of the WJ III, WJ IV, and WISC-V consistently indicate that very large portions of subtests' variance are apportioned to g and smaller (sometimes trivial) portions of unique subtest variance are associated with the factors representing broad abilities (i.e., Carroll's Stratum II) (Canivez & Watkins, 2016; Canivez et al., 2016, 2017; Dombrowski, 2013, 2014a, 2014b, 2015; Dombrowski et al., 2017, 2018a, 2018b; Dombrowski & Watkins, 2013). ω_H estimates are typically $> .75$, while ω_{HS} estimates are typically far $< .50$.³ Because the scores representing broad ability dimensions contain only small portions of variance independent of g , they are not adequately representing their intended CHC attributes. What makes this particularly remarkable is that these tests were designed to emphasize broad abilities, yet g accounts for the bulk of the variance. This is undeniably a threat to the veracity of the tests' measurement of CHC attributes, and perhaps the veracity of CHC itself as well.

Clinically, these results indicate that the reliability of the omnibus, full-scale scores on the WJ III, WJ IV, and WISC-V is satisfactory, but the same cannot be said for scores representing broad abilities. Most broad ability scores contain much more g variance than group factor variance indicating generally poor *unique* reliability of the broad ability scores. This impacts proper interpretation of the tests' broad ability composite scores due to the conflation of g variance and that unique to the broad ability factors. This conflation of variance renders such scores difficult to interpret because they represent a *mixture* of attributes (i.e., g + broad ability) instead of one particular attribute (Horn, 1989).

Another consequence of these observations relates to the internal consistency reliability estimates of scores included in test technical manuals. The internal consistency reliability methods used for broad ability scores likely violate major assumptions of the methods (Gignac & Watkins, 2013; Raykov, 1997; Watkins, 2017). As previously discussed, reliability estimates for the omnibus, full-scale scores consist of a large portion of true score variance, so traditional reliability estimates are similar to the ω_H estimates. This means that the confidence intervals used in score interpretation would also likely be similar. For the broad ability scores, however, the situation is quite different. Because ω_{HS} estimates are considerably lower than the traditional reliability estimates presented in test technical manuals, the standard error of measurement based on them would be considerably higher due to how much less *unique* true score variance is present. As a result, confidence intervals for the broad ability scores would be considerably larger when interpreting the *unique* broad abilities measurement – likely so large as to render them of little use.

2.4. g Factor Contribution

As illustrated in this article, intelligence tests typically capture large portions of variance due to g (Carroll's Stratum III), often the overwhelming majority of common variance. This means Carroll is likely correct in asserting the importance of accounting for g (Cucina & Howardson, 2017). Subsequently, the *de-emphasis* of g in favor of broad ability scores proffered by those promoting CHC-based score interpretations is likely misguided. This is apparent in manuals and interpretive guides for the WJ III, WJ IV, WISC-IV, and WISC-V where numerous CHC-based broad ability scores are the subject of interpretations, either on their own or in comparison with other broad ability scores (i.e., cognitive profile analysis). Similar de-emphasis of full-scale scores is present in neuropsychology (Hale & Fiorello, 2004) and cross-battery assessment approaches (Flanagan, Ortiz, & Alfonso, 2013, 2015). In fact, such a dialectic between emphasizing g versus broad abilities has been around since the very inception of clinical intelligence tests (Beaujean & Benson, 2019)

³The WISC-V Processing Speed (PS) score is an exception to this generalization.

The very notion that CHC-based broad ability scores somehow “inform” understanding about individuals’ particular cognitive abilities when they provide such little unique information apart from g seems to run counter to the replicated research in the extant literature. This is likely why many independent researchers recommend primary, if not exclusive, interpretation of the omnibus, full-scale scores; and if going beyond and interpreting broad ability scores, doing so with abundance of caution and awareness that there are risks of illusory correlation and confirmation bias in decision-making (e.g., Canivez et al., 2017, 2018; Dombrowski et al., 2017, 2018a, 2018b; Watkins, 2009).

2.5. Test Score Interpretation

Intelligence test technical and interpretation manuals (e.g., McGrew et al., 2014; Wechsler, 2014b) describe interpretation of broad ability scores as reflecting a specific attribute – usually one represented in a particular factor analysis (i.e., Verbal Comprehension). The absence of complete information about these scores (e.g., common and unique variance partitions) presents a situation where clinicians using these tests could interpret the score as though it represents measurement of only that specific ability. What is not made clear in test manuals is the fact that the broad ability scores provided are actually a mixture of g and broad ability variance. Thus, it is a fallacy to refer to, for example, the WISC-V VCI as assessing “verbal comprehension” when the ω_{HS} coefficient for the score indicates it contains < 25% true score variance *uniquely* attributed to “verbal comprehension” (see Canivez et al., 2017). Thus, the majority of true score variance contained in the VCI is not unique to verbal comprehension ability but to g !

Nowhere in CHC interpretative guides or publishers’ descriptions is the influence of g on broad ability scores acknowledged, and this substantially complicates what one might “infer” for individual clients, particularly because we do not know for particular individuals exactly how much of *their* performance on the VCI tasks, for example, is due to g and how much is due to verbal comprehension. Replicated research has consistently shown that portions of CHC broad ability variance are substantially smaller than g contributions; thus, attributions of such CHC-based scores representing broad abilities are misappropriated. If CHC is correct, one might expect that substantially larger portions of broad factor scores (Stratum II) ought to be based on that factor and not g . It may be that our present tests poorly measure anything supplemental to g , so it is difficult to determine exactly what their unique contributions are (Beaujean & Benson, 2018; Spearman & Wynn Jones, 1950)

2.6. Incremental Validity

Much consideration for the validity of intelligence test scores focuses on internal validity evidence (e.g., factor analysis). This provides only one necessary, yet not sufficient, condition for test validity support. Criterion-related validity evidence is important, and often paramount when the criterion has clinical or policy implications (Lubinski & Dawes, 1992).

In the case of intelligence tests, academic achievement is an important external variable and intelligence test scores ought to predict achievement well. Intelligence test technical manuals often include correlational studies of the intelligence test scores and academic test scores. It is common to see zero-order Pearson correlations between full-scale scores, broad ability scores, and subtest scores with a variety of scores from academic achievement tests (e.g., McGrew et al., 2014; Wechsler, 2014b). While this may be acceptable for the omnibus, full-scale composite score, it is not acceptable for broad ability because those scores conflate g variance and *unique* broad ability variance.

In order to identify *unique* contributions of broad ability scores to predicting academic achievement, it is crucial to first partial out the effects of g . Such incremental validity (Hunsley, 2003; Hunsley & Meyer, 2003) for broad ability scores is generally poor as the broad ability scores rarely account for *meaningful* portions of achievement apart from g (Canivez, 2013a; Canivez, Watkins, James, James, & Good, 2014; Glutting, Watkins, Konold, &

McDermott, 2006; Kranzler, Benson, & Floyd, 2015; McGill, 2015; McGill & Busse, 2015). For example, the WISC-IV *Technical and Interpretive Manual* (Wechsler, 2003) Table 5.15 (p. 69) presents zero-order Pearson correlations between the WISC-IV and WIAT-II from the standardization linking sample ($N = 550$) and the correlation between the WISC-IV FSIQ and WIAT-II Reading Composite score is .78, while the correlation between the WISC-IV VCI and WIAT-II Reading Composite score is .74. This gives the impression that the VCI is every bit as good as the FSIQ in predicting the Reading Composite score. What Glutting et al. (2006) showed in incremental validity assessment with the WISC-IV and WIAT-II standardization linking sample was that while the WISC-IV FSIQ and WIAT-II Reading Composite score was .78, the *part correlation* of VCI with the WIAT-II Reading Composite score was .04, meaning the *unique* contribution of VCI in accounting for WIAT-II Reading Composite score variance was .002 (.2%)! Thus, the zero-order correlation of the VCI with WIAT-II Reading Composite was primarily the result of the large portion of g variance within the VCI and not “verbal comprehension” at all. The reason for this is apparent given the results of structural validity studies presented earlier. The broad ability scores in tests such as the WJ III, WJ IV, WISC-IV, and WISC-V conflate g variance and *unique* variance due to the broad ability. So, when considering the ability of broad ability scores to account for academic achievement, usually it is the overwhelming portions of g variance in those scores that are contributing to the correlation, *not* the broad ability for which the score is named. None of the technical manuals of major intelligence tests provides such incremental validity statistics although this shortcoming has been reported (Canivez, 2010, 2014, 2017).

Some studies explicitly exploring CHC broad abilities relations with academic achievement may include some incremental validity assessment (e.g., Cormier, Bulut, et al., 2017; Cormier, McGrew, et al., 2017), while others (Cormier et al., 2016) do not. It is interesting that the Cormier, Bulut, et al. (2017) and Cormier, McGrew, et al. (2017) studies present one small paragraph reporting only the *increments* in R^2 provided by the WJ IV CHC cluster scores beyond g rather than presenting the amounts of mathematics or reading achievement variance accounted for by g *and* incremental improvement provided by the CHC broad ability scores. In contrast, several expansive paragraphs of results and discussion focused on the broad ability scores.

2.7. Ipsative and Pairwise Comparisons

Use of cognitive profile analysis (e.g., normative ipsative or pairwise score comparisons) to examine intracognitive variation within individuals is another frequently used approach proffered by CHC-based interpretive guides (e.g., Wechsler, 2014b). The history for such practices is long (Canivez, 2013b; Kamphaus, Winsor, Rowe, & Kim, 2012; McGill, Styck, Palomares, & Hass, 2016). Identifying such profiles to determine processing strengths and weaknesses (PSWs) among composite or subtest scores ignores the lack of evidence for reliability, validity, and diagnostic utility of such difference scores.

The resulting strengths and weaknesses obtained from cognitive profile analyses typically have minimal longitudinal stability – e.g., what is a strength at one time point may not be at a later point in time or could even be a weakness at a later time point (McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992; Watkins & Canivez, 2004; Watkins & Smith, 2013). Cronbach and Snow (1977) indicated that “any long-term recommendations as to a strategy for teaching a student would need to be based on aptitudes that are likely to remain stable for months, if not years” (p. 161), so for any patterns, profiles, or characteristics to have any clinical utility, an individual’s strengths and weaknesses must also be stable across time. Such stability in cognitive profiles has yet to be empirically demonstrated (e.g., Macmann & Barnett, 1997).

Inferences made about individual strengths or weaknesses are also problematic. The broad ability scores are thought of as estimates of some particular CHC ability, but as we previously discussed, frequently there is substantially more g variance within these scores than unique variance due to the broad abilities. This is a major problem for which appealing to CHC cannot overcome.

2.8. Diagnostic and Treatment Utility

In the assessment and treatment recommendations for specific learning disability (SLD), there have been methods proposed that are linked to PSWs of CHC attributes. A recent examination of State eligibility criteria for SLD (Maki, Floyd, & Roberson, 2015) found that 25 states did not permit PSW methods to be used, 14 specified PSW could be used, and 12 did not indicate whether or not PSW could be used. Of the 26 states that did not exclude PSW methods, 25 did not provide guidelines on *how* they should be used, while two states reportedly discussed use of CHC or alternative intelligence theories. Zirkel (2017) noted: “Although the special education literature clearly recognizes PSW as a distinct approach...state laws provide cryptic and unclear attention to this alternative for SLD identification, and the case law failures to address its contours and content” (p. 171). As McGill and Busse (2015) pointed out in their review, there are numerous problems and few examinations of diagnostic accuracy and treatment utility supporting the use of PSW. Those promoting PSW interpretations have failed to provide *a priori* examinations of the reliability and diagnostic utility of such indexes or their classification methods – something at odds with evidence-based assessment practices (Hunsley & Mash, 2011; McFall, 1991, 2000). PSW methods’ appeal to CHC is inadequate to support its use. In fact, the significant problems with CHC itself appear to reflect a poor foundation upon which to build any interpretive system. Diagnostic accuracy of assessment methods based on patterns or PSWs has met with disappointing results (Kranzler, Floyd, Benson, Zaboski, & Thibodaux, 2016a, 2016b; Miciak, Fletcher, Stuebing, Vaughn, & Tolar, 2014; Smith & Watkins, 2004; Watkins, 1999; Watkins, Kush, & Glutting, 1997). Likewise, PSW also provide poor guidelines for developing interventions (Burns et al., 2016). Because cognitive profile analyses do not contribute to diagnostic or treatment or intervention utility and clinicians are particularly susceptible to illusory correlations and confirmation bias in decision-making (Croskerry, 2003; Galanter & Patel, 2005; Glutting et al., 2003; Macmann & Barnett, 1997; Watkins, 2009), it is important to focus assessment activity on instruments and scores likely to have demonstrated utility (Norcross, Hogan, & Koocher, 2008; Youngstrom, 2013; Youngstrom et al., 2017). It has also been pointed out that administering additional tests or using longer tests is also not cost effective nor time effective (Glutting et al., 2003; Williams & Miciak, 2018).

2.9. CHC-Based Applications: XBA and PSW

Cross-battery assessment is an emblematic clinical application of CHC theory. In the days prior to the WJ III, no test appeared to adequately assess all of the supposedly important Horn-Cattell or Carroll identified broad abilities. This led to the recommendation that clinicians should “cross” batteries (Flanagan & McGrew, 1997; McGrew & Flanagan, 1998; Woodcock, 1990) to include more complete assessments of cognitive abilities beyond those offered in a singular test. McGrew and Flanagan focused interpretive emphasis on “*Gf-Gc pattern of abilities* of individuals that in turn can be related to important occupational and achievement outcomes and other human traits” but did not include *g* claiming it “has little practical relevance to cross-battery assessment and interpretation” (p. 14). The CHC taxonomy offered an organizational framework where clinicians could combine scores from multiple tests to more comprehensively assess those broad abilities. One such method presently in use is that of Flanagan, Ortiz, and Alfonso (2013, 2015). Another PSW method utilizing cross-battery methods is that proposed by Dehn and Szasz (2016), although not directly linked to CHC. Neither have provided sufficient evidence for the reliability, validity, or utility of their methods. Most of their support comes from stating they are based on CHC or linked to neuropsychology or “cognitive processes.” Given the lack of empirical evidence for PSW and XBA promotions, as well as past and recent negative results (e.g., Kranzler et al., 2016a, 2016b; McDermott et al., 1992; Watkins & Canivez, 2004), such methods should be discouraged in clinical practice until there is a sufficient evidence base supporting their use.

3. Conclusion

Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) require that test interpretation methods demonstrate empirical support and empirically supported test interpretation practices demand that such evidence be provided *a priori*. The adequacy of cognitive ability or intelligence tests in applied practice is primarily based on robust evidence for reliability, validity, and diagnostic utility of the scores produced. The overwhelming evidence for modern intelligence tests is that most of the subtests included appear to primarily measure *g* (Carroll's Stratum III) and only small portions of group factors that might be aligned with Carroll's Stratum II (or Horn-Cattell broad ability). Cucina and Howardson (2017) argued that their data supported Carroll and not Cattell-Horn and research reviewed in the present article also indicate dominance of *g* in tests that are supposedly reference instruments of CHC theory (WJ III and WJ IV) as well as the WISC-V. If the broad CHC abilities are to be the principal focus of interpretation and theoretical importance, they ought to be associated with substantially more unique variance than what is currently found in intelligence tests. The structural, incremental validity, reliability, and diagnostic utility problems for broad CHC abilities claimed to be measured by scores on the intelligence tests are poor and thus provide serious challenges to CHC. This, combined with McGrew's (2018a) own statement that he believed that *g* was a statistical artifact is consistent with Cattell and Horn's position, but not congruent with Carroll, and, as such, there appears to be little reason to include Carroll in CHC.

Those examining or applying intelligence tests in clinical practice who wish to deny the existence of *g* (following Horn, Cattell, and most CHC literature) must reconcile the abundance of evidence showing substantial covariance among the broad factors and the conflation of *g* and broad factor variance in broad factor/index scores. In contrast, those accepting the presence of *g* need only follow Carroll's 3S that *g* as well as a few orthogonal broad Stratum II abilities appear to be supported. This model does not require Cattell-Horn – something Carroll (2003) himself noted. A judge reviewing the evidence (i.e., incongruence of Cattell-Horn and Carroll's models, as well as their unfinished and longstanding debate, McGrew's [2018a, Schneider & McGrew, 2012] own comments that he disagrees with Carroll about *g*) would swiftly grant an annulment of the arranged and unhappy marriage of Horn and Cattell's E Gf-Gc (Horn & Blankson, 2005; Horn & Noll, 1997) and Carroll's 3S (Carroll, 1993, 2003). The decision would remove Carroll's "C" from the "CHC" moniker, and the models would be free to go their separate ways.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Gary L. Canivez  <http://orcid.org/0000-0002-5347-6534>

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Beaujean, A. A. (2015a). 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. (2016). Reproducing the Wechsler intelligence scale for children-fifth edition: Factor model results. *Journal of Psychoeducational Assessment*, 34, 404–408. doi:10.1177/0734282916642679
- Beaujean, A. A., & Benson, N. F. (2018, March). Theoretically-consistent cognitive ability test development and score interpretation. *Contemporary School Psychology*. Advance online publication. doi:10.1007/s40688-018-0182-1

- Beaujean, A. A., & Benson, N. F. (2019). The one and the many: Enduring legacies of Spearman and Thurstone on intelligence test score interpretation. *Applied Measurement in Education*.
- 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
- Burns, M. K., Peterson-Brown, S., Haegele, K., Rodriguez, M., Schmitt, B., Cooper, M., ... Hosp, J. (2016). Meta-analysis of academic interventions derived from neuropsychological data. *School Psychology Quarterly*, 31, 28–42. doi:10.1037/spq0000117
- Canivez, G. L. (2010). Test review of the Wechsler adult intelligence test – Fourth edition. In R. A. Spies, J. F. Carlson, & K. F. Geisinger (Eds.), *The eighteenth mental measurements yearbook* (pp. 684–688). Lincoln, NE: Buros Institute of Mental Measurements.
- Canivez, G. L. (2013a). Incremental validity of WAIS–IV factor index scores: Relationships with WIAT-II and WIAT-III subtest and composite scores. *Psychological Assessment*, 25, 484–495. doi:10.1037/a0032092
- Canivez, G. L. (2013b). Psychometric versus actuarial interpretation of intelligence and related aptitude batteries. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwane (Eds.), *The Oxford handbook of child psychological assessments* (pp. 84–112). New York, NY: Oxford University Press.
- Canivez, G. L. (2014). Test review of the Wechsler preschool and primary scale of intelligence – Fourth edition. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.), *The nineteenth mental measurements yearbook* (pp. 732–737). Lincoln, NE: Buros Center for Testing.
- Canivez, G. L. (2016). Bifactor modeling in construct validation of multifaceted tests: Implications for understanding multidimensional constructs and test interpretation. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advancements* (pp. 247–271). Göttingen, Germany: Hogrefe.
- Canivez, G. L., & Watkins, M. W. (2016). Review of the Wechsler Intelligence Scale for Children–Fifth Edition: Critique, commentary, and independent analyses. In A. S. Kaufman, S. E. Raiford, & D. L. Coalson (Eds.), *Intelligent testing with the WISC–V* (pp. 683–702). Hoboken, NJ: Wiley.
- 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.
- Canivez, G. L., Dombrowski, S. C., & Watkins, M. W. (2018). Factor structure of the WISC–V for four standardization age groups: Exploratory and hierarchical factor analyses with the 16 primary and secondary subtests. *Psychology in the Schools*, 55, 741–769. doi:10.1002/pits.22138
- Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2016). Factor structure of the Wechsler intelligence scale for children – Fifth edition: Exploratory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment*, 28, 975–986. doi:10.1037/pas0000238
- Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2017). Structural validity of the Wechsler intelligence scale for children – Fifth edition: Confirmatory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment*, 29, 458–472. doi:10.1037/pas0000358
- Canivez, G. L., Watkins, M. W., James, T., James, K., & Good, R. (2014). Incremental validity of WISC-IV^{UK} factor index scores with a referred Irish sample: Predicting performance on the WIAT-II^{UK}. *British Journal of Educational Psychology*, 84, 667–684. doi:10.1111/bjep.12056
- Carroll, J. B. (1993). *Human cognitive abilities*. Cambridge, England: Cambridge University Press.
- Carroll, J. B. (1995). On methodology in the study of cognitive abilities. *Multivariate Behavioral Research*, 30, 429–452. doi:10.1207/s15327906mbr3003_6
- Carroll, J. B. (1996). A three-stratum theory of intelligence: Spearman's contribution. In I. Dennis & P. Tapsfield (Eds.), *Human abilities: Their nature and measurement* (pp. 1–17). 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). New York, NY: Pergamon.
- Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. New York, NY: Elsevier.
- 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
- Cronbach, L., & Snow, R. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York, NY: Irvington.
- Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine*, 78, 775–780. doi:10.1097/00001888-200308000-00003

- Cucina, J. M., & Howardson, G. N. (2017). Woodcock-Johnson-III, Kaufman Adolescent and Adult Intelligence Test (KAIT), Kaufman Assessment Battery for Children (KABC), and Differential Ability Scales (DAS) Support Carroll but Not Cattell-Horn. *Psychological Assessment, 29*, 1001–1015. doi:10.1037/pas0000389
- Deary, I. J. (2001). Individual differences in cognition: British contributions over a century. *British Journal of Psychology, 92*, 217–237. doi:10.1348/000712601162040
- Dehn, M. J., & Szasz, C. (2016). *Psychological Processing Analyzer 5.6 (PPA)*. Sparta, WI: Schoolhouse Educational Services.
- Dombrowski, S. C. (2013). Investigating the structure of the WJ–III Cognitive at school age. *School Psychology Quarterly, 28*, 154–169. doi:10.1037/spq0000010
- Dombrowski, S. C. (2014a). Exploratory bifactor analysis of the WJ–III Cognitive in adulthood via the Schmid–Leiman procedure. *Journal of Psychoeducational Assessment, 32*, 330–341. doi:10.1177/0734282913508243
- Dombrowski, S. C. (2014b). Investigating the structure of the WJ–III Cognitive in early school age through two exploratory bifactor analysis procedures. *Journal of Psychoeducational Assessment, 32*, 483–494. doi:10.1177/0734282914530838
- 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., Canivez, G. L., & Watkins, M. W. (2018). Factor structure of the 10 WISC–V primary subtests across four standardization age groups. *Contemporary School Psychology, 22*, 90–104. doi:10.1007/s40688-017-0125-2
- 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., McGill, R. J., & Canivez, G. L. (2018a). An alternative conceptualization of the theoretical structure of the WJ IV Cognitive 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. (2018b). Exploratory and hierarchical factor analysis of the WJ IV Full Test battery. *School Psychology Quarterly, 33*, 235–250. doi:10.1037/spq0000221
- Dombrowski, S. C., & Watkins, M. W. (2013). Exploratory and higher order factor analysis of the WJ–III full test battery: A school aged analysis. *Psychological Assessment, 25*, 442–455. doi:10.1037/a0031335
- Elliott, C. D. (2007a). *Differential ability scales – Second edition*. San Antonio, TX: Harcourt Assessment, Inc.
- Elliott, C. D. (2007b). *Differential ability scales – Second edition: Introductory and technical handbook*. San Antonio, TX: Harcourt Assessment, Inc.
- Flanagan, D. P., & McGrew, K. S. (1997). A cross-battery approach to assessing and interpreting cognitive abilities: Narrowing the gap between practice and cognitive science. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 314–325). New York, NY: Guilford Press.
- Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). *Essentials of cross-battery assessment* (3rd ed.). Hoboken, NJ: Wiley.
- Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2015). *Cross-Battery Assessment Software System (X-BASS)*. Hoboken, NJ: John Wiley.
- Galanter, C. A., & Patel, V. L. (2005). Medical decision making: A selective review for child psychiatrists and psychologists. *Journal of Child Psychology and Psychiatry, 46*, 675–689. doi:10.1111/j.1469-7610.2005.01452.x
- Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model – Based reliability in the WAIS–IV. *Multivariate Behavioral Research, 48*, 639–662. doi:10.1080/00273171.2013.804398
- Glutting, J. J., Watkins, M. W., & Youngstrom, E. A. (2003). Multifactor and cross-battery assessments: Are they worth the effort? In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence aptitude, and achievement* (2nd ed., pp. 343–374). New York, NY: Guilford.
- Glutting, J. J., Watkins, M. W., Konold, T. R., & McDermott, P. A. (2006). Distinctions without a difference: The utility of observed versus latent factors from the WISC-IV in estimating reading and math achievement on the WIAT-II. *Journal of Special Education, 40*, 103–114. doi:10.1177/00224669060400020101
- Hale, J. B., & Fiorello, C. A. (2004). *School neuropsychology: A practitioner's handbook*. New York, NY: Guilford.
- Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*, 41–54. doi:10.1007/BF02287965
- Horn, J. L. (1989). Models of intelligence. In R. L. Linn (Ed.), *Intelligence, measurement, theory and public policy* (pp. 29–73). Urbana: University of Illinois Press.
- Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, & R. W. Woodcock (Eds.), *Woodcock-Johnson psycho-educational battery-revised technical manual* (pp. 197–232). Chicago, IL: Riverside.
- 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: Theories, tests, and issues* (pp. 53–91). New York, NY: Guilford.
- Horn, J. L., & Blankson, N. (2005). Foundations for better understanding of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 41–68). New York, NY: Guilford.

- Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligence. *Journal of Educational Psychology*, 57, 253–270. doi:10.1037/h0023816
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55. doi:10.1080/10705519909540118
- Hunsley, J. (2003). Introduction to the special section on incremental validity and utility in clinical assessment. *Psychological Assessment*, 15, 443–445. doi:10.1037/1040-3590.15.4.443
- Hunsley, J., & Mash, E. J. (2011). Evidence-based assessment. In D. H. Barlow (Ed.), *The Oxford handbook of clinical psychology* (pp. 76–97). New York, NY: Oxford University Press.
- Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment*, 15, 446–455. doi:10.1037/1040-3590.15.4.446
- Kamphaus, R. W., Winsor, A. P., Rowe, E. W., & Kim, S. (2012). A history of intelligence test interpretation. In D. P. Flanagan, L. Harrison, D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 56–70). New York, NY, US: Guilford Press.
- Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman assessment battery for children* (2nd ed.). Circle Pines, MN: American Guidance Service.
- Kranzler, J. H., Benson, N., & Floyd, R. G. (2015). Using estimated factor scores from a bifactor analysis to examine the unique effects of the latent variables measured by the WAIS-IV on academic achievement. *Psychological Assessment*, 27, 1402–1416. doi:10.1037/pas0000119
- Kranzler, J. H., Floyd, R. G., Benson, N., Zabolski, B., & Thibodaux, L. (2016a). Classification agreement analysis of cross-battery assessment in the identification of specific learning disorders in children and youth. *International Journal of School and Educational Psychology*. doi:10.1080/21683603.2016.1155515
- Kranzler, J. H., Floyd, R. G., Benson, N., Zabolski, B., & Thibodaux, L. (2016b). Cross-battery assessment pattern of strengths and weaknesses approach to the identification of specific learning disorders: Evidence-based practice or pseudoscience? *International Journal of School and Educational Psychology*. doi:10.1080/21683603.2016.1192855
- Littell, W. M. (1960). The Wechsler intelligence scale for children: Review of a decade of research. *Psychological Bulletin*, 57, 132–156. doi:10.1037/h0044513
- Lubinski, D., & Dawes, R. V. (1992). Aptitudes, skills, and proficiencies. In M. D. Dunnette & L. H. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3, pp. 1–59). Palo Alto, CA: Consulting Psychologists Press.
- 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
- Maki, K. E., Floyd, R. G., & Roberson, T. (2015). State learning disability eligibility criteria: A comprehensive review. *School Psychology Quarterly*, 30, 457–469. doi:10.1037/spq0000109
- 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. *The Journal of Special Education*, 25, 504–526. doi:10.1177/002246699202500407
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.
- McFall, R. M. (1991). Manifesto for a science of clinical psychology. *The Clinical Psychologist*, 44, 75–88.
- McFall, R. M. (2000). Elaborate reflections on a simple manifesto. *Applied and Preventive Psychology*, 9, 5–21. doi:10.1016/S0962-1849(05)80035-6
- McGill, R. J. (2015). Incremental criterion validity of the WJ-III COG clinical clusters. Marginal predictive effects beyond the general factor. *Canadian Journal of School Psychology*, 30, 51–63. doi:10.1177/0829573514560529
- McGill, R. J., & Busse, R. T. (2015). Incremental validity of the WJ III COG. Limited predictive effects beyond the GIA-E. *School Psychology Quarterly*, 30, 353–365. doi:10.1037/spq0000094
- McGill, R. J., & Dombrowski, S. C. (2019). Critically reflecting on the origins, evolution, and impact of the Cattell-Horn-Carroll (CHC) model. *Applied Measurement in Education*.
- McGill, R. J., Styck, K. M., Palomares, R. S., & Hass, M. R. (2016). Critical issues in specific learning disability identification: What we need to know about the PSW model. *Learning Disability Quarterly*, 39, 159–170. doi:10.1177/0731948715618504
- McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 136–181). New York, NY: Guilford.
- McGrew, K. S. (2014). *CHC Theory & Assessment: Best practices and what we’ve (I’ve) learned in the past 25 years*. Powerpoint slides. Retrieved from http://conference.esc13.net/assets/hci14/docs/McGrewAM_HCI14.pdf
- McGrew, K. S. (2018a, April 12). *Dr. Kevin McGrew and Updates to CHC Theory* [Video webcast]. Invited podcast presentation for School Psyched! Podcast presented 12 April 2018. Retrieved from <https://itunes.apple.com/us/podcast/episode-64-dr-kevin-mcgrew-and-updates-to-chc-theory/id1090744241?i=1000408728620&mt=2>
- McGrew, K. S. (2018b, May 18). *WJ IV norm-based and supplemental clinical test groupings for “intelligent” intelligence testing with the WJ IV (MindHub™ Pub. #3)*. St Joseph, MN: Institute for Applied Psychometrics.
- McGrew, K. S., & Flanagan, D. P. (1998). *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment*. Boston, MA: Allyn & Bacon.

- McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). *Technical manual: Woodcock-Johnson IV*. Rolling Meadows, IL: Riverside.
- McGrew, K. S., & Woodcock, R. W. (2001). *Technical manual: Woodcock-Johnson III*. Itasca, IL: Riverside Publishing.
- Miciak, J., Fletcher, J. M., Stuebing, K. K., Vaughn, S., & Tolar, T. D. (2014). Patterns of cognitive strengths and weaknesses: Identification rates, agreement, and validity for learning disabilities identification. *School Psychology Quarterly*, 29, 21–37. Retrieved from <http://dx.doi.org/10.1037/spq0000037>
- Norcross, J. C., Hogan, T. P., & Koocher, G. P. (2008). *Clinician's guide to evidence based practices: Mental health and the addictions*. London, UK: Oxford.
- Oh, H. J., Glutting, J. J., Watkins, M. W., Youngstrom, E. A., & McDermott, P. A. (2004). Correct interpretation of latent versus observed abilities: Implications from structural equation modeling applied to the WISC-III and WIAT linking sample. *Journal of Special Education*, 38, 159–173. doi:10.1177/00224669040380030301
- Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, 32, 329–353. doi:10.1207/s15327906mbr3204_2
- 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., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, 95, 129–140. doi:10.1080/00223891.2012.725437
- Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107, 358–367. doi:10.1037/0033-295X.107.2.358
- Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, 98, 223–237. doi:10.1080/00223891.2015.1089249f
- Roid, G. (2003a). *Stanford-Binet intelligence scales: Fifth edition*. Itasca, IL: Riverside Publishing.
- Roid, G. (2003b). *Stanford-Binet intelligence scales: Fifth edition, technical manual*. Itasca, IL: Riverside Publishing.
- 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, NY: Guilford Press.
- Schrank, F. A., McGrew, K. S., & Mather, N. (2014a). *Woodcock-Johnson IV*. Rolling Meadows, IL: Riverside.
- Schrank, F. A., McGrew, K. S., & Mather, N. (2014b). *Woodcock-Johnson IV tests of cognitive abilities*. Rolling Meadows, IL: Riverside.
- Silverstein, A. B. (1993). Type I, Type II, and other types of errors in pattern analysis. *Psychological Assessment*, 5, 72–74. doi:10.1037/1040-3590.5.1.72
- Smith, C. B., & Watkins, M. W. (2004). Diagnostic utility of the Bannatyne WISC-III pattern. *Learning Disabilities Research & Practice*, 19, 49–56. Retrieved from <https://doi.org/10.1111/j.1540-5826.2004.00089.x>
- Spearman, C. (1927). *The abilities of man*. New York, NY: Cambridge University Press.
- Spearman, C. E. (1931). Our need of some science in place of the word 'intelligence.' *Journal of Educational Psychology*, 22, 401–410. doi:10.1037/h0070599
- Spearman, C. E., & Wynn Jones, L. (1950). *Human ability: A continuation of "the abilities of man"*. London, UK: Macmillan.
- Strickland, T., Watkins, M. W., & Caterino, L. C. (2015). Structure of the Woodcock-Johnson III Cognitive Tests in a referral sample of elementary school students. *Psychological Assessment*, 27, 689–697. doi:10.1037/pas0000052
- Taub, G. E., & McGrew, K. S. (2014). The Woodcock-Johnson tests of cognitive abilities III's cognitive performance model: Empirical support for intermediate factors within CHC theory. *Journal of Psychoeducational Assessment*, 32, 187–201. doi:10.1177/0734282913504808
- Thurstone, L. L. (1938). *Primary mental abilities* (Psychometric Monographs 1). Chicago, IL: University of Chicago.
- Watkins, M. W. (1999). Diagnostic utility of WISC-III subtest variability among students with learning disabilities. *Canadian Journal of School Psychology*, 15, 11–20. doi:10.1177/082957359901500102
- 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. (2009). Errors in diagnostic decision making and clinical judgment. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (4th ed., pp. 210–229). New York, NY: Wiley.
- Watkins, M. W. (2017). The reliability of multidimensional neuropsychological measures: From alpha to omega. *The Clinical Neuropsychologist*, 31, 1113–1126. doi:10.1080/13854046.2017.1317364
- Watkins, M. W., & Canivez, G. L. (2004). Temporal stability of WISC-III subtest composite strengths and weaknesses. *Psychological Assessment*, 16, 133–138. doi:10.1037/1040-3590.16.2.133
- Watkins, M. W., Kush, J. C., & Glutting, J. J. (1997). Prevalence and diagnostic utility of the WISC-III SCAD profile among children with disabilities. *School Psychology Quarterly*, 12, 235–248. doi:10.1037/h0088960
- Watkins, M. W., & Smith, L. (2013). Long-term stability of the Wechsler Intelligence Scale for Children – Fourth Edition. *Psychological Assessment*, 25(477–483). doi:10.1037/a0031653

- Wechsler, D. (2003). *Wechsler intelligence scale for children-fourth edition technical and interpretive manual*. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (2014a). *Wechsler intelligence scale for children – Fifth edition*. San Antonio, TX: NCS Pearson.
- Wechsler, D. (2014b). *Wechsler intelligence scale for children – Fifth edition technical and interpretive manual*. San Antonio, TX: NCS Pearson.
- Williams, J., & Miciak, J. (2018). Adoption costs associated with processing strengths and weaknesses methods for learning disabilities identification. *School Psychology Forum: Research in Practice*, 12, 17–29.
- Woodcock, R. W. (1990). Theoretical foundations of the WJ-R measures of cognitive ability. *Journal of Psychoeducational Assessment*, 8, 231–258. doi:10.1177/073428299000800303
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III tests of cognitive abilities*. Rolling Meadows, IL: Riverside.
- Youngstrom, E. A. (2013). Future directions in psychological assessment: Combining evidence-based medicine innovations with psychology’s historical strengths to enhance utility. *Journal of Clinical Child & Adolescent Psychology*, 42, 139–159. doi:10.1080/15374416.2012.736358
- Youngstrom, E. A., Van Meter, A., Frazier, T. W., Hunsley, J., Prinstein, M. J., Ong, M.-L., & Youngstrom, J. K. (2017). Evidence-Based Assessment as an integrative model for applying psychological science to guide the voyage of treatment. *Clinical Psychology: Science and Practice*. doi:10.1111/cpsp.12207
- Zirkel, P. A. (2017). RTI and other approaches to SLD identification under the IDEA: A legal update. *Learning Disability Quarterly*, 40, 165–173. doi:10.1177/0731948717710778