Archives of Assessment Psychology, Vol. 6, No. 1, (49-79)
Printed in U.S.A.

Invalidating the Full Scale IQ Score in the Presence of Significant Factor Score Variability: Clinical Acumen or Clinical Illusion?

Ryan J. McGill, Ph.D., BCBA-D, NCSP

Abstract

Within the professional literature, it is frequently suggested that significant variability in lower-level factor and index scores on IQ tests renders the resulting FSIQ an inappropriate focus for clinical interpretation and diagnostic decision-making. To investigate the tenability of this popular interpretive heuristic, the present study examined the structural and predictive validity of the KABC-II for participants in the normative sample who were observed to have significant variability in their factor scores. Participants were children and adolescents, ages 7-18, (N = 2,025) drawn from the KABC-II/KTEA-II standardization sample. The sample was nationally stratified and proportional to U.S. census estimates for sex, ethnicity, geographic region, and parent education level. Using exploratory factor analysis and multiple factor extraction criteria, support for a five-factor extraction was obtained consistent with publisher theory. As recommended by Carroll (1993; 1995) hierarchical structure was explicated by sequentially partitioning variance appropriately to higher- and lower-order dimensions. Results showed the largest portions of total and common variance were accounted for by the second-order general factor with meaningful residual variance accounted for by Short-Term Memory at ages 7-12 and 13-18. As a result, the Fluid-Crystallized Index (FCI) accounted for large predictive effects across measures of academic achievement whereas the five first-order CHC factor scores consistently accounted for trivial proportions of incremental predictive variance beyond the FCI. Implications for clinical practice and the correct interpretation of the KABC-II and other related measurement instruments in the presence of significant scatter are discussed.

Introduction

As a result of advances in psychometric and neurocognitive theory, contemporary intelligence tests have been designed to appraise examinee performance at multiple levels (e.g., subtest scores, factor scores, global composites), providing examiners with the ability to make numerous inferences about the status of an individual's cognitive functioning (Canivez, 2013b). Accordingly, debates about the most useful procedures for interpreting the scores derived from

these measures are pervasive within the professional literature (Decker, Hale, & Flanagan, 2013; Watkins, 2000). Whereas some scholars contend that the global ability score (i.e., full scale IQ [FSIQ]) is the most parsimonious and valid predictor of important life outcomes such as achievement and occupational attainment (e.g., Dombrowki & Gischlar, 2014; Canivez, 2013b; Gottfredson, 1997; Schmidt & Hunter, 2004), others suggest that the profile of lower-order factor and index scores provides users with more useful information than the FSIQ for more focal diagnostic decision-making and treatment planning (Feifer et al., 2014; Fiorello et al., 2007; Hale & Fiorello, 2001).

Issues with Cognitive Profile Analysis

Primary interpretation of factor and index score profiles for diagnostic decision-making has long been advocated in the technical literature despite suggestions that these approaches are new and revolutionary (e.g., Flanagan, Ortiz, Alfonso, & Dynda, 2006; Fiorello, Hale, & Wycoff, 2012). Over 70 years ago, Rapaport et al. (1945) proposed an interpretive framework that provided clinicians with a step-by-step process for analyzing intra-individual cognitive strengths and weaknesses based upon the belief that variations in cognitive test performance serve as potential evidence for the presence of a variety of clinical disorders and a multitude of related approaches have been subsequently developed (e.g., Kaufman, 1994; Naglieri, 2000; Priftera & Dersh, 1993).

As a result, the trend among publishers has been to create longer test batteries that provide users with an ever increasing number of composite indices (Glutting, Watkins, & Youngstrom, 2003). As a consequence, a considerable amount of time and resources are expended by psychologists to administer and interpret the wealth of information provided by these instruments (Yates & Taub, 2003). This investment is based upon the assumption that the additional information provided beyond the more global FSIQ is clinically useful. To illustrate, Pfeiffer, Reddy, Kletzel, Schmelzer, and Boyer (2000) surveyed 354 nationally certified school psychologists regarding their use and perceptions of profile analysis and reported approximately 70% of respondents believed that the information obtained from profile analysis was clinically meaningful and 89% of respondents declared that they used profile analysis routinely when making diagnostic decisions. More recently, Decker, Hale, and Flanagan (2013) suggested that profile analysis has become even more prevalent in clinical and school psychology due to the popularity of cross-battery (XBA; Flanagan, Ortiz, & Alfonso, 2013) and other related interpretive approaches.

Whereas the psychometric shortcomings of subtest-level profile analysis have long been known (Macmann & Barnett, 1997; McDermott et al., 1992; McDermott, Fantuzzo, & Glutting, 1990), there is a countering body of evidence that brings into question the primary interpretation of intelligence tests at the factor score level. Structural validity investigations have revealed conflicting factor structures from those reported in the technical manuals of contemporary cognitive measures (e.g., Canivez, 2008; Canivez & Watkins, 2010; Dombrowski, Canivez, Watkins, & Beaujean, 2015; Dombrowki, 2013), suggesting that these instruments may be overfactored (Frazier & Youngstrom, 2007). Additionally, the long-term stability and diagnostic utility of these indices has also been found wanting (Watkins, 2000; Watkins & Smith, 2013). Most recently, McDermott, Watkins, and Rhoad (2014) found that a significant amount of factor-level variability across long-term retest intervals was attributable to variables that had nothing to do with individual differences (e.g., assessor bias), posing a significant threat to inferences made from cognitive profile data at any one point in time.

Additionally, the emergence of bifactor modeling in the psychometric literature raises questions about the accuracy of procedures (e.g., coefficient alpha) used to estimate the internal consistency of factor scores on cognitive measures. As an example, Canivez (2014) examined the WISC-IV with a referred sample and found that the factor-level scores were inherently multidimensional (i.e., composed of non-trivial proportions of construct irrelevant variance attributable to the higher-order general factor). According to Beaujean, Parkin, and Parker (2014), multidimensionality is not the problem per se; the problem occurs when an interpretation of individual cognitive abilities and their related composites "fails to recognize that Stratum II factors derived from higher-order models are not totally independent of *g*'s influence" (p. 800). As Horn (1991) cautioned long ago, attempting to disentangle the different features of cognition is akin to "slicing smoke." Whereas it may be possible for practitioners to account for general factor effects when interpreting primarily at the factor-level, contemporary profile analysis models have yet to provide a mechanism for doing so (McGill, Styck, Palomares, & Hass, 2015).

In sum, these measurement concerns threaten confident interpretation of factor-level profiles as diagnostic decisions based on data obtained from measures that have questionable psychometric properties will be hopelessly flawed (Dawes, Faust, & Meehl, 1989). As Fletcher et al. (2013) argued, "It is ironic that methods of this sort [profile analysis] continue to be proposed when the basic psychometric issues are well understood and have been documented for many years" (p. 40).

Utility and Stability of the Global FSIQ Score

In contrast to factor-level scores, psychometric support for global FSIQ score, and other related indices, is strong and includes the highest internal consistency estimates, short- and long-term stability estimates, and predictive validity coefficients (Canivez, 2013b). As a consequence, practitioners have been encouraged to focus most of their interpretive weight on the FSIQ, and to interpret information provided by lower-order factor and index scores cautiously, if at all, due to the aforementioned psychometric concerns at that level of measurement (Glutting, Watkins, & Youngstrom, 2003; Kranzler & Floyd, 2013). Nevertheless, questions about the relevance of the FSIQ when significant variability is observed between its constituent factor and index scores have long been raised by researchers. That is, "is there a statistical or clinical point where FSIQ 'fractures' into more meaningful parts, and is no longer a valid measure of general mental ability nor clinically useful for assisting with differential diagnosis or program planning" (Beal et al., 2016, p. 66)?

While Drozdick, Wahlstrom, Zhu, and Weiss (2012) suggest that extreme score discrepancies do not automatically invalidate the FSIQ, recommendations to eschew reporting and/or interpreting the FSIQ in the presence of significant interfactor variability are ubiquitous and long standing within the professional literature. In fact, due to the popularity of this interpretive heuristic (heretofore referred to as the 'variability hypothesis'), it may be argued that the variability hypothesis serves as a proverbial *lingua franca* for clinical IQ test interpretation across applied psychological disciplines (e.g., clinical and school psychology). To wit, in the popular *Handbook of Psychological Assessment*, Groth-Marnat (2009) noted that "Examiners can interpret the more global measures (FSIQ) with greater meaning, usefulness and certainty if there is not a high degree of difference amongst the index scores or other groupings…With increasing differences, the purity of the global measures becomes contaminated" (p. 140). Hale and Fiorello (2004) were even more definitive in their recommendations to school psychologists,

encouraging practitioners to "just say no" to interpretation of the FSIQ score when variability is observed at any level of the measurement instrument. To wit, "you should *never* [emphasis added] report an IQ score whenever there is significant subtest or factor variability…and any interpretation of that IQ score would be considered inappropriate" (p. 100). Not surprisingly, the interpretive manuals for many contemporary cognitive tests provide users with detailed procedures for accounting for the variability hypothesis in their clinical interpretations of FSIQ scores and other related global composites.

As an example, the Technical and Interpretive Manuals for the latest iterations of the Wechsler Scales (Wechsler, 2008; 2014) encourage users to interpret scores in a stepwise fashion beginning with the FSIQ and then proceeding to the factor scores after examining the consistency of the scores contained within those indicators. That is, for the FSIQ to be interpreted, the variability between the lower-order factor scores must not exceed *a priori* thresholds, denoting varying degrees of statistical and clinical significance (e.g., 15-20 standard score points). If meaningful variability is observed, users are encouraged to forego clinical interpretation of the FSIQ and focus all of their interpretive weight on the profile of obtained factor scores. While related procedures on rival IQ tests vary, they all stress that the putative absence of factor score variability is a necessary condition for the FSIQ to be considered meaningful and/or interpretable (Reschly, Myers, & Hartel, 2002).

According to Marley and Levin (2011), prescriptive statements such as these in education and psychology are rarely justified and require adherence to high standards of empirical evidence. Relatedly, Haynes, Smith, and Hunsley (2011) stress that interpretive procedures for psychological assessments, including those recommended for accounting for the variability hypothesis (e.g., Hale & Fiorello, 2004), must be supported with evidence obtained from appropriate validity studies. However, no validity evidence has been provided in the Technical and Interpretive Manuals for the Wechsler Scales or other rival measurement instruments to support these interpretive procedures, which is in direct conflict with validity standards contained in the most recent edition of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

Results of Previous Studies Examining the Variability Hypothesis

In what was termed the first direct examination of the effect of score variability on the predictive validity of the FSIQ score, Watkins, Glutting, and Lei (2007) found that the WISC-III FSIQ score remained a more robust predictor of academic achievement when compared to the lower-order factor scores and that there was no interaction effect for profile group based upon the observed level of score variability with a mixed normative/clinical sample. As a result, the authors challenged the practice of discounting the FSIQ score as a predictor of academic achievement when factor scores significantly vary. These results were later replicated on the Differential Ability Scales (DAS; Kotz, Watkins, & McDermott, 2008) and in a later study predicting long-term achievement outcomes with the WISC-III (Freberg, Vandiver, Watkins, & Canivez, 2008).

In terms of construct validity, Fiorello and colleagues (2002) utilized regression communality analysis to examine the constitution of the WISC-III FSIQ in a sample of typical children with flat (n = 707) and variable cognitive profiles (n = 166). Whereas it was found that the FSIQ communality for the flat subsample was primarily composed of unique variance (i.e., *g*), the FSIQ communality for the variable subsample was composed mostly of shared variance,

suggesting attenuation of the general factor due to score variability. Based upon these results, they suggested that the FSIQ does not represent global ability for individuals with significant levels of scatter. These results were later replicated with a sample of children with learning disabilities in mathematics on the DAS-II (Hale et al., 2008), a finding that they attributed to the discordant cognitive profiles frequently observed within those samples. However, the use of communality analysis as a method for higher-order variance partitioning is controversial. In 2007, a special issue of *Applied Neuropsychology* was commissioned by the journal editor to debate the use of such methods. In a commentary, Dana and Dawes (2007) were critical of the conclusions reached by the Fiorello and Hale research group and questioned why a more appropriate technique (e.g., factor analysis) was not utilized to examine the structure of intellectual functioning. Schneider (2008) later criticized the use of communality analysis for explanatory purposes, likening it to the use of an "Ouija Board," suggesting that it was an inappropriate procedure for making inferences about latent structure. In their response to these criticisms, Hale et al. (2007) argued that a *g*-factor was only plausible if manifest variables were observed to load on a single latent dimension, an extreme position at odds with the factor analytic literature (e.g., Carroll, 1993, 1995; Watkins, 2006).

In a more direct appraisal of the effect of index score scatter on the structural validity of the FSIQ, Daniel (2007) utilized exploratory factor analysis to examine the amount of variance explained by the first un-rotated factor with simulation data designed to mimic varying degrees of factor variability. In general, it was found that the observed FSIQ remained an equally valid summary of global cognitive ability for groups with variable and flat cognitive profiles. As an explanation for the findings, Daniel (2007) concluded that the influence of score variabilities tends to counteract, rendering their net effect on the global composite to be trivial. It should be noted that hierarchical structure was not explicated; therefore, the conjoint effects of variability on higher-order and lower-order scores (e.g., first-order factors and subtests) remain unexamined.

Failure to Examine Potential Effects on Hierarchical Structure of Variables

According to Carroll (2003), all cognitive measures are composed of reliable variance that is attributable to a second-order general factor, reliable variance that is attributable to first-order group factors, and error variance. Because of this, Carroll argued that variance from the second-order factor must be extracted first to residualize the first-order factors, leaving them orthogonal to the second-order dimension. Thus, variability associated with a second-order factor is accounted for before interpreting variability associated with first-order factors, resulting in variance being apportioned correctly to higher- and lower-order dimensions. To accomplish this task, Carroll (1993, 1995) recommended second-order exploratory factor analysis (EFA) of first-order factor correlations followed by a Schmid-Leiman transformation (Schmid & Leiman, 1957). The Schmid-Leiman technique allows for the orthogonalization of second-order variance from first-order factors. According to Carroll (1995):

> 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 variance decomposition procedures described above are a potentially useful vehicle for examining the tenability of the variability hypothesis as they provide direct estimates of the proportion of *g* variance contained within higher- and lower-order scores. However, these procedures have yet to be employed for such purposes, suggesting that our understanding of the variability hypothesis is presently incomplete.

Limitations of Previous Research

While previous studies examining the predictive validity of scores in the presence of significant factor score variability (e.g., Freberg, Vandiver, Watkins, & Canivez, 2008; Watkins, Glutting, & Lei, 2007; Kotz, Watkins, & McDermott, 2008) have consistently found that the predictive effects of the FSIQ fail to be attenuated, these studies have largely employed various iterations of the Wechsler Scales and other related measurement instruments that fail to incorporate modern theories of cognitive abilities, such as the Cattell-Horn-Carroll model (CHC; Schneider & McGrew, 2012), as part of their foundation. A more significant limitation has been the dearth of investigations designed to examine the potential impact of variability on the latent structure of measurement instruments. Whereas, in the only study designed specifically for these purposes (Daniel, 2007), it was found that a higher-order factor could still be plausibly extracted in the presence of significant factor score variability, the effects of scatter on the latent composition of higher- and lower-order scores was not fully explored. Additionally, the EFA conducted by Daniel utilized simulated data to examine the effects of variability on higher-order structure. While there is nothing wrong with simulations per se, examination of these effects using normative data would be more instructive for informing clinical practice, as these samples serve as the foundation for many of the quantitative and qualitative inferences that clinicians make with the data obtained from their administrations of IQ tests (Glutting, McDermott, Watkins, & Kush, 1997).

Purpose and Goals of the Current Study

To address these gaps in the literature, the present investigation examined the effect of significant factor variability on the structural and incremental validity of the Kaufman Assessment Battery for Children-Second Edition (KABC-II; Kaufman & Kaufman, 2004a) using the hierarchical EFA procedures recommended by Carroll (1993, 1995) and supplemented with hierarchical multiple regression. Since users of the KABC-II are encouraged to invalidate the global composite if significant factor scatter is observed amongst the lower-order indicators (i.e., Kaufman, Lichtenberger, Fletcher-Janzen, & Kaufman, 2005; Lichtenberger, Sotelo-Dynega, & Kaufman, 2009), there is a critical need to provide practitioners with information regarding the psychometric validity of this interpretive heuristic with potential applications for applied practice. Accordingly, the purpose of the present study is to: a) determine whether hierarchical structure remains plausible for participants from the KABC-II normative sample who present with significant levels of score variability, b) examine the proportions of KABC-II subtest variance attributed to the second-order general dimension (e.g., *g*) and to the first-order latent dimensions using the hierarchical exploratory techniques described by Carroll (1993, 1995), and c) to assess the incremental validity of the factor scores in accounting for meaningful portions of Kaufman Test of Educational Achievement-Second Edition (KTEA-II; Kaufman & Kaufman, 2004c) achievement variance when significant scatter is observed.

Similar procedures have been utilized successfully to examine the structural and predictive validity of factor-level scores on many contemporary cognitive measures (e.g.,

Canivez, 2011, 2013a; Canivez & Watkins, 2010; Dombrowski , 2013; McGill & Busse, 2015), which facilitates comparison of results obtained from the present investigation to these other studies.

In addition to being the first study to apply these procedures to investigate the tenability of the variability hypothesis, it is believed that examining the potential effects of variability on KABC-II scores will be instructive for generalizing the results obtained from previous research to an instrument that utilizes CHC theory as part of its theoretical and interpretive foundation. The current study is an extension of previous research and will potentially provide practitioners and researchers with important information about the correct interpretation of the KABC-II and other related measurement instruments when significant factor scatter is observed.

Participants

Participants were children and adolescents ages 7-0 to 18-11 (N = 2,025) drawn from the KABC-II/KTEA-II standardization sample. Demographic characteristics are provided in detail in the KABC-II manual (Kaufman & Kaufman, 2004b). The standardization sample was obtained using stratified proportional sampling across demographic variables of age, sex, race/ethnicity, parent educational level, and geographic region. Examination of the tables in the manual revealed a close correspondence to the 2001 U. S. census estimates across the stratification variables. The present sample was selected on the basis that it corresponded to the age ranges at which the publisher proposed measurement model could be fully specified as well as the fact that it permitted analyses of relationships between cognitive variables across a clinically relevant age span (e.g., primary and secondary school-age).

Measurement Instruments

Kaufman Assessment Battery for Children-Second Edition. The KABC-II (Kaufman & Kaufman, 2004a) measures the processing and cognitive abilities of children and adolescents between the ages of 3 years and 18 years. The KABC-II utilizes a dual-theoretical foundation featuring the CHC psychometric model of broad and narrow abilities (Schneider & McGrew, 2012), and Luria's neuropsychological theory of cognitive processing (Luria, 1966). Although model choice is up to the examiner, the KABC-II manual encourages users to interpret the KABC-II primarily from the CHC perspective. The CHC interpretive model for ages 7-18 (outlined graphically in Figure 1) features 10 core subtests, which combine to yield five first-order factor scale scores (Short-Term Memory, Long-Term Storage and Retrieval, Visual Processing, Fluid Reasoning, and Crystallized Ability), and a second-order Fluid Crystallized Index (FCI) that is thought to represent psychometric *g*. Each CHC factor scale is composed of two subtest measures, and the FCI is derived from a linear combination of the 10 core subtests that compose the constituent factor scores. It should be noted that from ages 3-6, the KABC-II utilizes different subtest measures and not all latent dimensions of the school-age CHC models are replicated. Additionally, the subtests that compose the Visual Processing factor differ from ages 7-12 and 13-18 (see Figure 1).

All factor and composite variables on the KABC-II are expressed as standard scores with a mean of 100 and a standard deviation of 15. The total norming sample (N = 3,025) is nationally representative based upon 2001 U.S. census estimates. Extensive normative and psychometric data can be found in the KABC-II manual (Kaufman & Kaufman, 2004b). Mean internal consistency estimates for the included ages in this study ranged from .88 to .93 for the factor

scores and .77 to .93 for the subtests. The mean internal consistency estimate for the FCI was .97. Validity evidence is provided in several forms in the KABC-II manual and independent reviews are available (e.g., Bain & Gray, 2008; Braden & Ouzts, 2005).

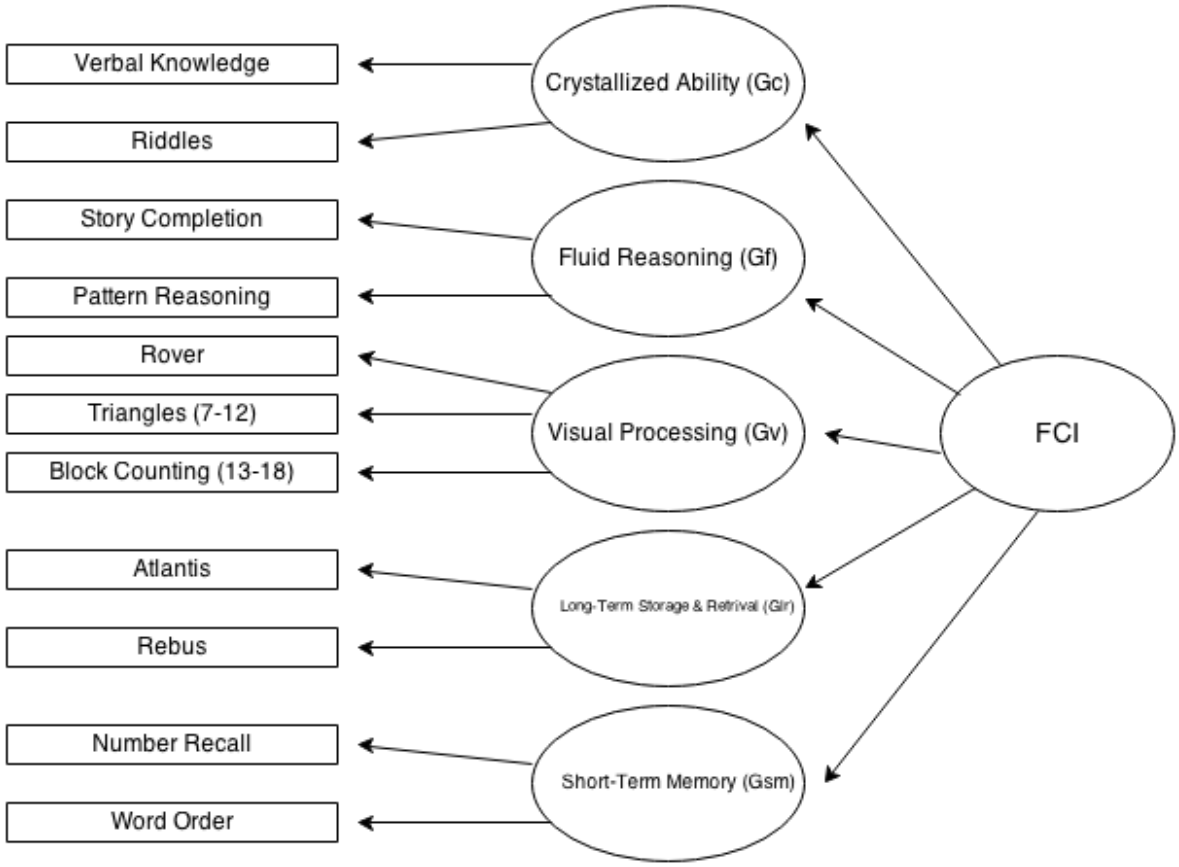

*Figure 1***.** Indirect hierarchical Cattell-Horn-Carroll (CHC) interpretive model for the KABC-II. Adapted with permission from the KABC-II Technical Manual (Kaufman & Kaufman, 2004b).

In terms of clinical interpretation, the manual suggests that users should interpret scores in a stepwise fashion beginning with the FCI and then proceeding to more specific measures (e.g., indexes and subtests). While the test authors suggest that interpretation of the global composite is "almost always secondary in importance to fluctuations within the scale profile" (p. 43), specific guidance on when invalidation of the FCI would be warranted is not provided. However, in supplemental interpretive resources, procedures for invalidating the FCI have been more explicit (Kaufman, Lichtenberger, Fletcher-Janzen, & Kaufman, 2005; Lichtenberger, Sotelo-Dynega, & Kaufman, 2009). To wit, "If the variability between indexes on the KABC-II [difference between highest and lowest score] is 23 points or greater, then the meaningfulness of the global score is diminished. In such cases we encourage examiners to focus interpretation on the profile of scale indexes and to not interpret the global score" (Kaufman, Lichtenberger, Flecther-Janzen, & Kaufman, 2005, p. 86). Although it is suggested that this *rule* is straightforward and easy to remember it was not supported with any empirical evidence. In contrast, the validity of this approach was supported on the basis that it corresponded to

interpretive procedures advocated in a similar resource for WISC-IV interpretation (Flanagan & Kaufman, 2004).

Kaufman Test of Educational Achievement-Second Edition. The KTEA-II is a comprehensive academic assessment battery designed to measure four academic domains: Reading, Mathematics, Written Language, and Oral Language. The KTEA-II is comprised of 14 subtests that combine to yield 4 domain composites and a total achievement composite score. All scores are expressed as standard scores with a mean of 100 and a standard deviation of 15. Mean internal consistency estimates for the included ages in this study ranged from .93 to .96 for composite scores that were assessed. Additional technical information for the KTEA-II can be found in the KTEA-II manual (Kaufman & Kaufman, 2004d).

Procedure

According to the KABC-II manual, all normative participants in the KABC-II dataset were administered measures from the KABC-II and the KTEA-II by trained examiners under the direct supervision of a standardization project member. Additionally, each examinee was given the measurement instruments in counterbalanced order.

Data Analyses

Data analyses for the present study occurred in several steps. First, to accord with the recommend procedures for invalidating the FCI composite in clinical practice (i.e., Kaufman, Lichtenberger, Fletcher-Janzen, & Kaufman, 2005), pairwise comparisons for all KABC-II factor score permutations were created in the normative dataset to select participants who presented with a 23 point or higher discrepancy between their highest and lowest CHC factor standard scores. Selected cases were then screened using the duplicate case analysis function in SPSS version 23. To further insure the integrity of the sample, 25% of the selected cases were randomly selected, using a random number generator, for verification of meeting inclusion criteria. Next, cases were filtered by age to permit separate structural and predictive validity analyses at ages 7-12 and 13-18.

Principal axis EFA (Fabrigar & Wegner, 2012; Fabrigar, Wegener, MacCallum, & Strahan, 1999) was used to analyze the reliable common variance from each of the two KABC-II variability sample correlation matrices representing the 10 core subtests that combine to form the CHC interpretive model using SPSS version 23 for Windows. As recommended by Gorsuch (1983), multiple criteria for determining the number of factors to retain were examined. These procedures included the visual scree test (Cattell, 1966), Horn's parallel analysis (HPA; Horn, 1965), and minimum average partials (MAP; Velicer, 1976). While the scree test was used to visually determine the optimum number of factors to retain, it is a subjective methodology. As recommended by Frazier and Youngstrom (2007), HPA and MAP were also included as they potentially protect against the threat of overfactoring in EFA. HPA and MAP analyses were conducted via the R statistical programming language (R Development Core Team, 2016) using the Psych package (Revelle, 2012).

For hierarchical exploratory analyses, the current study limited iterations in first-order principle axis factor extraction to two in estimating final communality estimates. According to Gorsuch (2003), limiting iterations to two provides an optimal balance between sampling and measurement error in estimating communality. Each correlation matrix for the two KABC-II samples with significant variability was subjected to a first-run EFA (principal axis extraction),

followed by a promax (oblique) rotation (k = 4; Gorsuch, 2003). The resulting first-order factors were orthogonalized in a second-run by removing variance associated with the higher-order general factor via the Schmid and Leiman (SL; 1957) procedure using the MacOrtho program (Watkins, 2004). According to Schmid and Leiman (1957), this transforms "an oblique solution containing a hierarchy of higher order factors into an orthogonal solution which not only preserves the desired interpretation characteristics of the oblique solution, but also discloses the hierarchical structuring of the variables" (p. 53). As per Child (2006), salient factor loading coefficients were defined as those≥ .30. Consistent with Carroll (1993), theoretically consistent non-trivial residual loading coefficients (e.g., .20 to .29) were considered to be *aligned*.

Next, omega-hierarchical and omega-subscale (Reise, 2012) were estimated as model-based reliability estimates of the latent factors (Gignac & Watkins, 2013). Chen, Hayes, Carver, Laurenceau, and Zhang (2012) stressed that "for multidimensional constructs, the alpha coefficient is complexly determined, and McDonald's omega-hierarchical (ω_h; 1999) provides a better estimate for the composite score and thus should be used" (p. 228). Omega-subscale (ω_s) is the model-based reliability estimate of a group factor with all other group *and* general factors removed (Reise, 2012). Omega estimates (ω_h and ω_s) may be obtained from Orthogonalized SL EFA solutions and were produced using the *Omega* program (Watkins, 2013). Albeit subjective, omega coefficients should at a minimum exceed .50, but .75 would be preferred (Reise, 2012; Reise, Bonifay, & Haviland, 2013).

Incremental validity (Hunsley, 2003) was assessed through hierarchical multiple regression analyses with the KTEA-II composite scores for Reading, Math, and Written Language serving as dependent variables. The KABC-II FCI was entered into the first block of the regression equation and the CHC factor scores were entered both jointly and individually into the second block. The change in predicted achievement variance (R^2) produced by CHC factor score(s) in the second block indicated their incremental prediction of achievement beyond the FCI composite score. The results were interpreted using the resulting R^2 statistic as an effect size. Guidelines for interpreting R^2 as an effect size are found in Cohen (1988); they are "small," .01; "medium," .09; and "large," .25.

Results

Analysis of the KABC-II/KTEA-II normative data indicated that 1,209 participants ages 7-18 (59% of the total normative sample for that age range) presented with the requisite level of CHC factor score profile variability for invalidating the FCI global composite (Kaufman, Lichtenberger, Fletcher-Janzen, & Kaufman, 2005; Lichtenberger, Sotelo-Dynega, & Kaufman, 2009). The distribution of significant variability was relatively invariant across the 7-12 (60%) and 13-18 (58%) age ranges. KABC-II subtest intercorrelation matrices and descriptive statistics for normative participants with significant profile variability are reported in Table 1. Score distributions were relatively normal with .39 the largest univariate skew and .43 the largest index of univariate kurtosis (Fabrigar et al., 1999).

Factor-Extraction Criteria

Whereas parallel analysis (Horn, 1965) and visual scree suggested that five factors be retained for both age ranges, the MAP (Velicer, 1976) criterion recommended retention of one factor for both age ranges. Given that it is better to over factor than under factor (Wood, Tataryn, & Gorsuch, 1996), five factors were extracted for both age groups to accord with the theoretical structure delineated in the KABC-II manual (Kaufman & Kaufman, 2004b).

Table 1

Intercorrelation Matrices and Descriptive Statistics for Kaufman Assessment Battery for Children-Second Edition (KABC-II) CHC Model Core Subtests for Ages 7-18 (*N* = 1,209)

Ages 7-12 (*n* = 694)

Variable	RI	VK	NR	WO	AT	RL	RO	TR	SC	PR
Riddles (RI)	-									
Verbal Knowledge (VK)	.73	-								
Number Recall (NR)	.24	.20	-							
Word Order (WO)	.25	.20	.57	-						
Atlantis (AT)	.36	.29	.10	.10	-					
Rebus (RL)	.42	.37	.14	.17	.56	-				
Rover (RO)	.29	.22	.11	.16	.15	.26	-			
Triangles (TR)	.37	.35	.15	.14	.22	.27	.45	-		
Story Completion (SC)	.42	.46	.12	.17	.22	.33	.21	.35	-	
Pattern Reasoning (PR)	.44	.48	.18	.24	.27	.36	.35	.43	.49	-

	RI	VK	NR	WO	AT	RL	RO	TR	SC	PR
M	10.22	10.18	10.39	10.11	10.20	10.39	10.24	10.06	10.24	10.11
SD	3.18	3.13	3.01	3.00	3.25	3.32	3.01	3.00	3.15	3.02
SK	0.04	0.00	0.00	0.39	-0.26	-0.14	-0.04	-0.03	0.02	-0.06
K	- 0.41	-0.19	-0.12	0.16	0.36	-0.04	-0.22	-0.26	-0.22	-0.27

Ages 13-18 (*n* = 515)

Variable	RI	VK	NR	WO	AT	RL	RO	BC	SC	PR
Riddles (RI)	-									
Verbal Knowledge (VK)	.78	-								
Number Recall (NR)	.22	.21	-							
Word Order (WO)	.30	.24	.59	-						
Atlantis (AT)	.35	.31	.12	.20	-					
Rebus (RL)	.41	.39	.12	.22	.56	-				
Rover (RO)	.19	.21	.08	.21	.13	.29	-			
Block Counting (BC)	.28	.29	.11	.20	.11	.27	.45	-		
Story Completion (SC)	.40	.40	.05	.11	.13	.37	.22	.26	-	
Pattern Reasoning (PR)	.40	.42	.21	.22	.23	.42	.36	.40	.48	-

	RI	VK	NR	WO	AT	RL	RO	TR	SC	PR
M	10.44	10.12	10.18	10.12	9.97	10.37	10.21	10.17	10.30	10.30
SD	3.15	3.12	3.07	3.05	3.34	2.99	3.24	3.18	3.10	3.06
SK	-0.18	0.11	-0.08	0.11	-0.02	-0.33	-0.11	-0.02	0.03	-0.17
K	-0.39	-0.21	-0.29	-0.29	-0.19	-0.01	-0.22	-0.25	-0.28	0.43

Note. Values rounded to nearest hundredth. All coefficients were statistically significant ($p < .01$, two-tailed). *SK* = Skewness. *K* = Kurtosis.

Exploratory Factor Analyses: Ages 7-12

The results of Bartlett's Test of Sphericity indicated that the correlation matrix was not random χ^2 (45) = 2,179.18, $p < .001$, and the Kaiser-Meyer-Olkin Measure of Sampling Adequacy coefficient of .80 was well above the minimum standard for conducting factor analysis (Kaiser, 1974). Communality estimates ranged from .384 (Rover) to .710 (Riddles). On the basis

of these values, it was determined that the correlation matrix was appropriate for the EFA procedures that were employed (Kaiser, 1974). Table 2 presents results from extracting five KABC-II factors with promax (k = 4) rotation. The *g* loadings (factor structure coefficients from first un-rotated factor) ranged from .296 (Number Recall) to .809 (Riddles) with the majority of the subtests in the fair to good range based on Kaufman's (1994) criteria (≥ .70 = good, .50 –.69 = fair, < .50 = poor). All KABC-II subtests were saliently and properly associated with their theoretical factor demonstrating desirable simple structure. The moderate to high factor correlations presented in Table 2 (.33 to .71) imply a higher-order (general intelligence) factor structure requiring explication (Carroll, 1998; Gorsuch, 1983; Thompson, 2004).

Hierarchical SL EFA Model Ages 7-12

The five first-order oblique EFA factor solution (Table 2) was transformed with the SL orthogonalization procedure. Results for the higher-order factor analysis of five first-order

Table 2

KABC-II Exploratory Factor Analysis Results for Standardization Sample Ages 7-12 with Significant Score Variability (N = 694) Five Oblique Factor Solution

		Factor Pattern Coefficients					Factor Structure Coefficients				
Subtest	*g*[a]	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
Riddles	.809	**.807**	.071	.035	.073	-.090	.838	.529	.344	.503	.589
Verbal Knowledge	.788	**.809**	-.051	-.027	-.060	.139	.836	.452	.285	.452	.637
Number Recall	.296	.038	-.009	**.718**	-.013	-.048	.256	.162	.709	.206	.209
Word Order	.322	-.037	-.002	**.722**	.002	.048	.262	.186	.725	.251	.267
Atlantis	.446	.012	**.726**	-.023	-.027	-.041	.374	.692	.142	.298	.337
Rebus	.547	-.009	**.700**	.016	.022	.068	.458	.747	.220	.413	.465
Rover	.401	-.025	.004	.008	**.687**	-.088	.298	.279	.195	.615	.382
Triangles	.520	.038	-.023	-.031	**.604**	.101	.420	.341	.210	.675	.530
Story Completion	.584	.106	.005	-.025	-.065	**.633**	.516	.373	.209	.434	.658
Pattern Reasoning	.635	.015	.019	.042	.165	**.559**	.532	.424	.297	.590	.711

	F1	F2	F3	F4	F5
Eigenvalue	3.78	1.38	1.09	0.95	0.70
Variance (%)	37.83	13.79	10.86	9.49	6.95

Factor Correlation Matrix

F1	1.00				
F2	.57	1.00			
F3	.37	.25	1.00		
F4	.55	.49	.33	1.00	
F5	.71	.55	.34	.70	1.00

Note. Salient factor pattern loadings (e.g., ≥ .30) are denoted in bold. *g* = general intelligence.
[a]Factor structure coefficients from first un-rotated factor (*g*-loadings) are correlations between subtest and general factor (Jensen, 1980).

KABC-II factors are presented in Table 3. All subtests were properly associated (higher residual variance) with their theoretically proposed factor after removing *g* variance. The hierarchical *g* factor accounted for 28.3% of the total variance and 58.5% of the common variance.

The general factor also accounted for between 7.4% (Number Recall) and 47.6% (Riddles) of individual subtest variability. At the first-order level, the Crystallized Ability factor accounted for an additional 4.5% of the total variance and 8.6% of the common variance, the Long-Term Retrieval factor accounted for an additional 5.9% of the total variance and 11.1% of the common variance, the Short-Term Memory factor accounted for an additional 8.6% of the total variance and 16.3% of the common variance, the Visual Processing factor accounted for an additional 3.8% of the total variance and 7.2% of the common variance, and the Fluid Reasoning factor accounted for an additional 1.7% of the total variance and 3.2% of the common variance. The general and group factors combined to measure 53% of the variance in KABC-II scores resulting in 47% unique variance (combination of specific & error variance). Subtest specificity (reliable variance unique to the individual measures) ranged from .15 to .41.

Table 3

Sources of Variance in the Kaufman Assessment Battery for Children-Second Edition (KABC-II) Normative Sample for Ages 7-12 with Significant Score Variability (N = 694) According to an Orthogonalized Higher-Order Factor Model

	General		F1: Gc		F2: Glr		F3: Gsm		F4: Gv		F5: Gf			
Subtest	*b*	Var	*b*	Var	*b*	Var	*b*	Var	*b*	Var	*b*	Var	h^2	u^2
RI	.690	.476	**.473**	**.224**	.054	.003	.032	.001	.048	.002	-.043	.002	.708	.292
VK	.688	.473	**.474**	**.225**	-.039	.002	-.025	.001	-.040	.002	.066	.004	.706	.294
NR	.272	.074	.022	.000	-.007	.000	**.653**	**.426**	-.009	.000	-.023	.001	.502	.498
WO	.313	.098	-.022	.000	-.002	.000	**.657**	**.432**	.001	.000	.023	.001	.531	.469
AT	.418	.175	.007	.000	**.550**	**.303**	-.021	.000	-.018	.000	-.019	.000	.478	.522
RL	.533	.284	-.005	.000	**.530**	**.281**	.015	.000	.015	.000	.032	.001	.566	.434
RO	.423	.179	-.015	.000	.003	.000	.007	.000	**.454**	**.206**	-.042	.002	.387	.613
TR	.545	.297	.022	.000	-.017	.000	-.028	.001	**.399**	**.159**	.048	.002	.460	.540
SC	.588	.346	.062	.004	.004	.000	-.023	.001	-.043	.002	**.299**	**.089**	.441	.559
PR	.658	.433	.009	.000	.014	.000	.038	.001	.109	.012	**.264**	**.070**	.516	.484
%TV	28.3		4.5		5.9		8.6		3.8		1.7		53.0	47.0
%CV	58.5		8.6		11.1		16.3		7.2		3.2		100.0	
	ω_h = .733		ω_s = .264		ω_s = .385		ω_s = .567		ω_s = .258		ω_s = .108			

Note. *b* = standardized loading of subtest on factor, Var = variance (b^2) explained in the subtest, h^2 = communality, u^2 = uniqueness, Gc = Crystallized Ability, Gsm = Short-Term Memory, Glr = Long-Term Storage and Retrieval, Gv = Visual Processing, Gf = Fluid Reasoning, RI = Riddles, VK = Verbal Knowledge, NR = Number Recall, WO = Word Order, AT = Atlantis, RL = Rebus, RO = Rover, TR = Triangles, SC = Story Completion, PR = Pattern Reasoning, TV = Total variance, CV = Common variance. Bold denotes loadings on theoretically assigned latent dimension. ω_h = Omega hierarchical, ω_s = Omega subscale.

Omega-hierarchical and omega-subscale coefficients were estimated based on the SL results in Table 3. The ω_h coefficient for general intelligence (.733) was high and sufficient for scale interpretation. The ω_s coefficients for the five specific KABC-II group factors were considerably lower ranging from .108 to .567. Whereas the coefficient obtained for Short-Term Memory (.567) was sufficient for interpretation, the four remaining CHC group factors likely possess too little unique true score variance for clinical interpretation (Reise, 2012; Reise et al., 2013).

Exploratory Factor Analyses: Ages 13-18

The results of Bartlett's Test of Sphericity indicated that the correlation matrix was not random χ^2 (45) = 1,697.94, p < .001, and the Kaiser-Meyer-Olkin Measure of Sampling Adequacy coefficient of .76 was well above the minimum standard for conducting factor analysis (Kaiser, 1974). Communality estimates ranged from .409 (Rover) to .762 (Riddles). On the basis of these values, it was determined that the correlation matrix was appropriate for the EFA procedures that were employed (Kaiser, 1974). Table 4 presents results from extracting five KABC-II factors with promax (k = 4) rotation. The *g* loadings (factor structure coefficients from first un-rotated factor) ranged from .317 (Number Recall) to .745 (Riddles) with the majority of the subtests in the fair to good range based on Kaufman's (1994) criteria (≥ .70 = good, .50 –.69 = fair, < .50 = poor). All KABC-II subtests were saliently and properly associated with their theoretical factor demonstrating desirable simple structure. The moderate to high factor correlations presented in Table 2 (.30 to .68) imply a higher-order (general intelligence) factor structure requiring explication (Carroll, 1998; Gorsuch, 1983; Thompson, 2004).

Table 4

KABC-II Exploratory Factor Analysis Results for Standardization Sample Ages 13-18 with Significant Score Variability (N = 515) Five Oblique Factor Solution

	Factor Pattern Coefficients						Factor Structure Coefficients				
Subtest	*g*[a]	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
Riddles	.745	**.831**	.026	.056	-.006	.005	.871	.358	.509	.382	.557
Verbal Knowledge	.726	**.828**	-.018	-.013	.018	.056	.856	.314	.464	.393	.572
Number Recall	.317	-.010	**.764**	-.051	-.097	.052	.245	.724	.173	.179	.156
Word Order	.416	.016	**.720**	.053	.091	-.078	.307	.754	.286	.320	.217
Atlantis	.449	.067	.006	**.755**	-.049	-.148	.362	.215	.689	.214	.287
Rebus	.634	-.053	-.022	**.675**	.045	.196	.441	.241	.770	.447	.568
Rover	.420	-.053	-.016	.034	**.678**	-.042	.219	.201	.276	.636	.403
Block Counting	.471	.075	-.019	-.072	**.660**	.001	.309	.218	.254	.654	.453
Story Completion	.527	.092	-.057	-.052	-.067	**.682**	.439	.121	.333	.394	.648
Pattern Reasoning	.652	-.013	.088	.007	.161	**.579**	.453	.298	.425	.585	.708

	F1	F2	F3	F4	F5
Eigenvalue	3.69	1.39	1.20	0.99	0.75
Variance (%)	36.86	13.85	12.02	9.88	7.47

Factor Correlation Matrix

F1	1.00				
F2	.38	1.00			
F3	.54	.32	1.00		
F4	.42	.35	.44	1.00	
F5	.62	.30	.56	.68	1.00

Note. Salient factor pattern loadings (e.g., ≥ .30) are denoted in bold. *g* = general intelligence.
[a]Factor structure coefficients from first un-rotated factor (*g*-loadings) are correlations between subtest and general factor (Jensen, 1980).

Hierarchical SL EFA Model Ages 13-18

The five first-order oblique EFA factor solution (Table 4) was transformed with the SL orthogonalization procedure. Results for the higher-order factor analysis of five first-order KABC-II factors are presented in Table 5. All subtests were properly associated (higher residual variance) with their theoretically proposed factor after removing *g* variance. The hierarchical *g* factor accounted for 26.7% of the total variance and 48.4% of the common variance.

The general factor also accounted for between 7.2% (Number Recall) and 42.4% (Riddles) of individual subtest variability. At the first-order level, the Crystallized Ability factor accounted for an additional 6.7% of the total variance and 12.1% of the common variance, the Short-Term Memory factor accounted for an additional 9% of the total variance and 16.4% of the common variance, the Long-Term Retrieval factor accounted for an additional 5.6% of the total variance and 10.2% of the common variance, the Visual Processing factor accounted for an additional 4.8% of the total variance and 8.7% of the common variance, and the Fluid Reasoning factor accounted for an additional 2.3% of the total variance and 4.2% of the common variance. The general and group factors combined to measure 55.2% of the variance in KABC-II scores resulting in 44.8% unique variance (combination of specific & error variance). Subtest specificity (reliable variance unique to the individual measures) ranged from .11 to .43.

Table 5

Sources of Variance in the Kaufman Assessment Battery for Children-Second Edition (KABC-II) Normative Sample for Ages 13-18 with Significant Score Variability (N = 515) According to an Orthogonalized Higher-Order Factor Model

	General		F1: Gc		F2: Gsm		F3: Glr		F4: Gv		F5: Gf			
Subtest	*b*	Var	*b*	Var	*b*	Var	*b*	Var	*b*	Var	*b*	Var	h^2	u^2
RI	.651	.424	**.573**	**.328**	.023	.001	.041	.002	-.004	.000	.003	.000	.754	.246
VK	.643	.413	**.571**	**.326**	-.016	.000	-.010	.000	.013	.000	.029	.001	.741	.259
NR	.268	.072	-.007	.000	**.687**	**.472**	-.037	.001	-.069	.005	.027	.001	.551	.449
WO	.360	.130	.011	.000	**.648**	**.420**	.039	.002	.065	.004	-.040	.002	.557	.443
AT	.403	.162	.046	.002	.005	.000	**.554**	**.307**	-.035	.001	-.077	.006	.479	.521
RL	.610	.372	-.037	.001	-.020	.000	**.496**	**.246**	.032	.001	.101	.010	.631	.369
RO	.419	.176	-.037	.001	-.014	.000	.025	.001	**.482**	**.232**	-.022	.000	.411	.589
BC	.463	.214	.052	.003	-.017	.000	-.053	.003	**.469**	**.220**	.001	.000	.440	.560
SC	.543	.295	.063	.004	-.051	.003	-.038	.001	-.048	.002	**.353**	**.125**	.430	.570
PR	.643	.413	-.009	.000	.079	.006	.005	.000	.114	.013	**.300**	**.090**	.523	.477
% TV	26.7		6.7		9.0		5.6		4.8		2.3		55.2	44.8
%CV	48.4		12.1		16.4		10.2		8.7		4.2		100.0	
	ω_h = .713		ω_s = .375		ω_s = .578		ω_s = .363		ω_s = .318		ω_s = .147			

Note. *b* = standardized loading of subtest on factor, Var = variance (b^2) explained in the subtest, h^2 = communality, u^2 = uniqueness, Gc = Crystallized Ability, Gsm = Short-Term Memory, Glr = Long-Term Storage and Retrieval, Gv = Visual Processing, Gf = Fluid Reasoning, RI = Riddles, VK = Verbal Knowledge, NR = Number Recall, WO = Word Order, AT = Atlantis, RL = Rebus, RO = Rover, BC = Block Counting, SC = Story Completion, PR = Pattern Reasoning, TV = Total variance, CV = Common variance. Bold denotes loadings on theoretically assigned latent dimension. ω_h = Omega hierarchical, ω_s = Omega subscale.

Omega-hierarchical and omega-subscale coefficients were estimated based on the SL results in Table 5. The ω_h coefficient for general intelligence (.713) was high and sufficient for

scale interpretation. The ω_s coefficients for the five specific KABC-II group factors were considerably lower ranging from .147 to .578. Whereas the coefficient obtained for Short-Term Memory was sufficient for interpretation, the remaining four group factors likely possess too little unique true score variance for clinical interpretation (Reise, 2012; Reise et al., 2013).

Incremental Predictive Validity

Hierarchical multiple regression analyses are presented in Table 6 and illustrate the proportions of KTEA-II achievement variance accounted for by the KABC-II FCI in the first block followed by the unique incremental variance accounted for by the lower-order CHC factor scores in the second block (after partialing out FCI variance). The change in R^2 from the second block provided the estimate of the incremental prediction provided by the CHC factors both jointly and individually. Using Cohen's (1988) guidelines for interpreting R^2 as an effect size, the FCI accounted for large and statistically significant effects across the KTEA-II reading, math, and written language indicators for both age groups. Although the incremental predictive contributions of the CHC factor scores across all achievement variables were statistically significant ($p < .05$), effect size estimates for these effects were consistently small with the exception of the moderate effects that were observed at ages 13-18 in predicting reading beyond the FCI. Additional achievement variance accounted for by individual CHC factors ranged from 2% to 10% (*Mdn* = 3.5%). With only the Crystallized Ability factor contributing anything beyond trivial effects (9%) in the reading model for ages 13-18.

Table 6

Incremental Contribution of Observed KABC-II CHC Factor Scores in Predicting KTEA-II Scores beyond the FCI for Normative Sample Participants with Significant Score Variability Ages 7-18 (N = 1,209)

Ages 7-12 (*n* = 694)

	Reading Composite			Math Composite			Written Language		
	R^2	Δ R^2	Δ (%)[b]	R^2	Δ R^2	Δ (%)[b]	R^2	Δ R^2	Δ (%)[b]
FCI	.51*	-	51%	.45*	-	45%	.39*	-	39%
CHC Factor Scores (*df* = 5)[a]	.55	.04*	4%	.47	.02*	2%	.42	.03*	3%
Crystallized Ability	.54	.03*	3%	.46	.01*	1%	.40	.01*	1%
Fluid Reasoning	.52	.01*	1%	.45	.00	0%	.40	.01*	1%
Visual Processing	.53	.01*	1%	.45	.00	0%	.41	.01*	1%
Long-Term Storage & Retrieval	.51	.00	0%	.46	.01*	1%	.40	.01*	1%
Short-Term Memory	.51	.00	0%	.45	.00	0%	.39	.00	0%

Ages 13-18 (*n* = 515)

	Reading Composite			Math Composite			Written Language		
	R^2	Δ R^2	Δ (%)[b]	R^2	Δ R^2	Δ (%)[b]	R^2	Δ R^2	Δ (%)[b]
FCI	.52*	-	52%	.49*	-	49%	.41*	-	41%
CHC Factor Scores (*df* = 5)[a]	.62	.10*	10%	.52	.03*	3%	.47	.06*	6%
Crystallized Ability	.61	.09*	9%	.50	.01*	1%	.45	.04*	4%
Fluid Reasoning	.52	.00	0%	.50	.01*	1%	.41	.00	0%
Visual Processing	.55	.03*	3%	.49	.00	0%	.45	.04*	4%
Long-Term Storage & Retrieval	.52	.00	0%	.49	.00	0%	.42	.00	0%
Short-Term Memory	.52	.00	0%	.50	.01*	1%	.41	.00	0%

Note. FCI = Fluid-Crystallized Index score. CHC = Cattell-Horn-Carroll factor scores. All coefficients rounded to nearest hundredth, may not equate due to rounding.

[a]Degrees of freedom reflects controlling for the effects of the FCI.
[b]Represents proportion of variance accounted for by variables at their entry point into regression equation. $R^2/\Delta R^2$ values multiplied by 100.
* $p < .05$.

Discussion

Within the assessment literature, it is routinely suggested that significant variability among factor and index scores renders the global IQ score an invalid measure of overall ability (e.g., Flanagan, Ortiz, & Alfonso, 2013; Lezak, Howieson, Bigler, & Tranel, 2012). In those cases, psychologists are encouraged to disregard the FSIQ and focus their interpretive weight on the profile of lower-order scores when engaging in diagnostic decision-making (e.g., Courville, Coalson, Kaufman, & Raiford, 2016). Although the techniques to account for the effects of scatter on the FSIQ vary across measures, they all assume that meaningful levels of scatter fracture the global composite, rendering it inconsequential for interpretive foci (Reschly, Myers, & Hartel, 2002). As recommended by methodologists (e.g., Dana & Dawes, 2007; Daniel, 2007), the present study sought to examine the tenability of this assumption via multiple techniques (i.e., Wasserman & Bracken, 2013) for appraising the latent structure and predictive validity of KABC-II (Kaufman & Kaufman, 2004a) scores, using participants in the KABC-II/KTEA-II normative sample that presented with significant levels of scatter in their CHC-based factor-level scores.

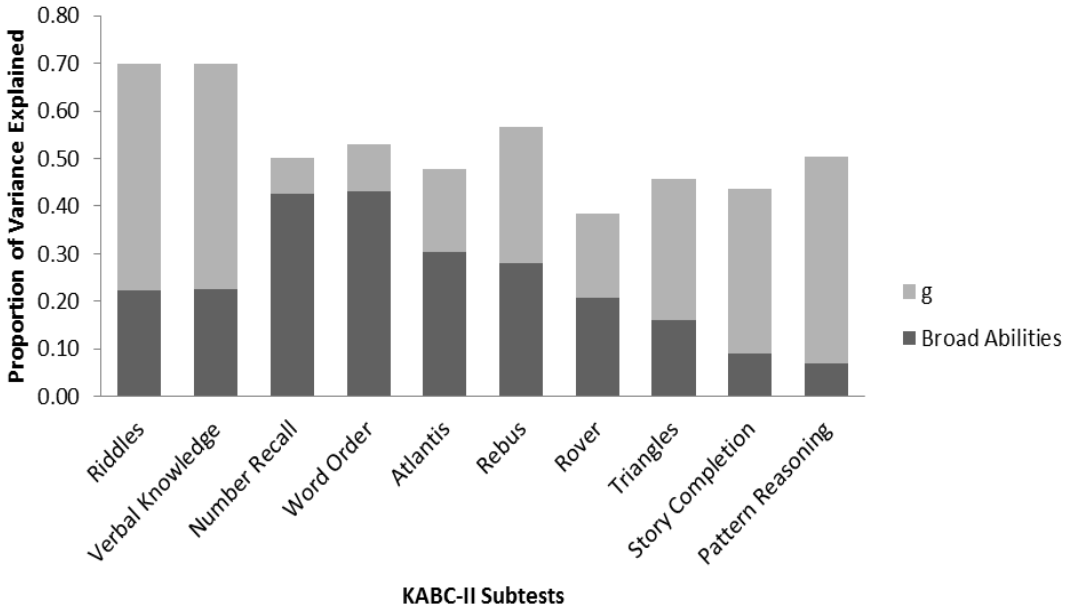

Figure 2. Proportion of KABC-II CHC model core subtests for ages 7-12 explained by the general factor and by the broad abilities.

EFA results using multiple factor extraction criteria supported the extraction of five factors for both age groups using the ten subtest CHC model configuration. While all of the KABC-II subtests were properly associated with their theoretically proposed latent first-order group factors (Kaufman & Kaufman, 2004b) for all age groups, the extraction of five factors resulted in moderate to highly correlated first-order dimensions, suggesting the presence of a

second-order general factor. Gignac (2007) has encouraged researchers to always perform orthogonalization procedures when examining higher-order model solutions. Thus, in order to better understand the underlying structure of the KABC-II in the presence of scatter, prescribed procedures (e.g., Carroll, 1993; 1995; Schmid & Leiman, 1957) were utilized in order to correctly apportion subtest variance appropriately to higher- and lower-order dimensions.

The application of the Schmid and Leiman (1957) transformation to the KABC-II school-age samples demonstrated that variance for each of the core subtests that comprise the CHC interpretive model can be decomposed into multiple components. Examination of variance apportions to the hierarchical *g* factor and the group factors found substantially greater total and common variance associated with the hierarchical *g* factor. In fact, the general factor explained more than twice the amount of the total variance of any single domain specific factor despite attenuation in positive manifold evidenced in the subtest intercorrelations reported in Table 1.

Whereas the combination of *g* and uniqueness outweighed the contributions made by the five first-order CHC factors (see Tables 3 and 5), meaningful common variance were accounted for by Short-Term Memory (Gsm) at ages 7-12 (16%) and 13-18 (16.4%), suggesting that additional consideration of Gsm may provide users with useful information as it relates to individual performance beyond *g* when significant levels of scatter are observed. Although this finding is not inconsequential, it should be noted that similar results have been found when appraising the latent structure of instruments designed to measure some of the same neurocognitive constructs as the KABC-II (e.g., Canivez, 2011; Strickland, Watkins, & Caterino, 2015). Nevertheless, as a consequence of the weaker portions of true score variance in the remaining group factors (as estimated by ω_s coefficients) there appears to be little variance apart from *g* in these factor scores to warrant clinical interpretation (Reise, 2012; Reise et al., 2013), or if interpreted, done with extreme caution.

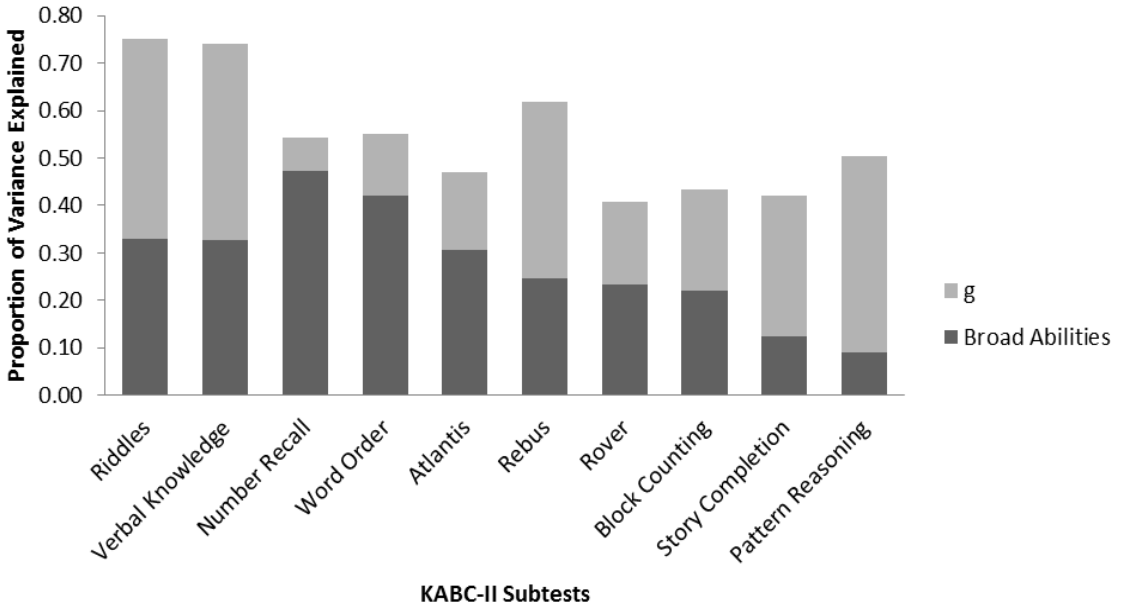

Figure 3. Proportion of KABC-II CHC model core subtests for ages 13-18 explained by the general factor and by the broad abilities.

The present results illustrate well that even in the presence of significant levels of scatter, the hierarchical structuring of cognitive dimensions (e.g., Carroll, 1993; Schneider & McGrew, 2012) was not altered, findings consistent with those obtained by Daniel (2007) in a simulation study using WISC-IV index parameters. More importantly, scatter did not render the second-order *g*-factor that was produced for the KABC-II inviolate as its effects must be accounted for at every level of the measurement instrument (see Figures 2 and 3). While general factor attenuation was clearly evident, especially with respect to its effects on the Short-Term Memory and Long-Term Retrieval dimensions, the total and common variance accounted for by *g* in the present study were relatively similar to estimates obtained by McGill and Spurgin (2015) when they utilized these same EFA procedures to fit the theoretical model suggested by the test publisher to the correlation matrices produced for entire KABC-II normative sample. However, the present results divulge from those obtained by McGill and Spurgin (2015) as the effects of scatter appeared to help clarify and support the tenability of the five-factor CHC model suggested by the test publisher. Whereas, forcing the theoretical model to the entire normative sample resulted in theoretically inconsistent subtest migration, weak subtest loadings, and impermissible factors, desired simple structure was obtained across both age groups in the current study.

As noted by Nelson and Canivez (2012), validity studies of the internal structure of intelligence tests, although necessary, are insufficient in informing higher-order versus lower-order interpretations. As a consequence, examination of incremental validity is particularly informative for determining the appropriate interpretation of KABC-II scores in the presence of significant score variability. Hierarchical multiple regression results were less ambiguous as the FCI accounted for large effects in all of the KTEA-II indicators that were assessed at both age groups. Conversely, the incremental variance (after controlling for the effects of the FCI) accounted for by the CHC factor scores was consistently trivial, with only a moderate effect observed in predicting reading beyond the FCI at ages 13-18. Curiously, the FCI accounted for more, with the CHC factors accounting for less, math variance at ages 7-12 when compared to the same model coefficients obtained by McGill (2015) with the total KABC-II normative sample. In sum, these results are consistent with the previous literature indicating that scatter does not render the global composite an invalid predictor of achievement (e.g., Freberg, Vandiver, Watkins, & Canivez, 2008; Watkins, Glutting, & Lei, 2007; Kotz, Watkins, & McDermott, 2008).

Base Rates and the Barnum Effect

Kaufman et al. (2005) suggest that the meaningfulness of the global composite on the KABC-II is diminished if variability amongst the first-order indexes meets or exceeds a critical value of 23 points. While this is based upon the belief that significant scatter is unique and rare, Watkins (2003) suggested that this and other related interpretive heuristics fails to account for variation that is common in the population. As the present results demonstrate, over half of the participants in the school-age KABC-II normative sample presented with significant levels of scatter, despite the use of a relatively conservative critical level. Other measurement instruments (e.g., Wechsler Scales) suggest much lower thresholds (e.g., 15 points) for considering whether the scatter that is observed between obtained factor-level scores is clinically significant and thus corrupting the FSIQ.

Long ago, Paul Meehl (1956) identified what he termed a *Barnum effect* in clinical assessment that occurs when psychologists generate hypotheses from markers that have high base rates of occurrence within the population. According to Wiggins (1988), this can result in diagnostic impressions that are true of virtually all people of the type that is under consideration. In the presence of such high base rates; clinicians must be mindful of the threat of *illusory correlation*—the false belief that two variables are related (Chapman & Chapman, 1967) when attempting to utilize the presence of significant scatter as potential diagnostic sign. While additional research is needed to determine if these base rates generalize to other measurement instruments, clinicians should bear in mind that significant variation is endemic in the population and thus must be interpreted cautiously (Flanagan, Alfonso, & Mascolo, 2011). As a safeguard, it is frequently recommended in the literature that clinicians seek to confirm or disconfirm hypotheses with additional sources of information, including the potential administration of additional cognitive measures (e.g., Decker, Hale, & Flanagan, 2013; Fiorello, Hale, & Wycoff, 2012). Though, due to the litany of decision errors that are possible when attempting to aggregate multiple sources of data, Faust (1990) warned clinicians long ago that the "common belief in the capacity [of clinicians] to perform complex configural analysis and data integration might thus be appropriately described as a shared professional myth" (p. 478).

Limitations and Future Directions

This study is not without limitations that should be considered when interpreting the results. The most important limitation of the present study is the use of an archived standardization sample. Although the sample was relatively large and nationally representative, additional research is needed to determine if these results generalize to more focal clinical populations such as those with specific learning disabilities and other related neurocognitive impairments as well as other measurement instruments.

Additionally, the use of confirmatory factor analytic (CFA) techniques to investigate the tenability of rival measurement models (e.g., correlated factors, bifactor) and/or the invariance of measurement models across groups with significant and non-significant levels of scatter for the KABC-II and other related measurement instruments would benefit assessment psychologists. Such examinations would permit clinicians to be able consider the convergence or divergence of CFA and EFA results and the potential implications for clinical interpretation. Despite the substantive changes to the KABC-II (including the specification of a new theoretical model), EFA was eschewed in favor of a constrained CFA to validate internal structure. However evidence for improper solutions (e.g., Heywood cases) were found in the resulting school-age validation models (see Figure 8.2, Kaufman & Kaufman, 2004b). Unfortunately, the results of an independent EFA investigation (McGill & Spurgin, 2015) did not support the theoretical model suggested by the test publisher or the fit of more parsimonious models. Carroll (1995) recommended that "a confirmatory analysis of a dataset should not be published without an accompanying statement or report on one or more appropriate exploratory analyses" (p. 437). As a result, EFA was utilized in the present study to provide results that would be of potential benefit to subsequent CFA investigations. Although EFA and CFA are considered to be complimentary procedures, Gorsuch (1983, 2003) noted that they provide answers to different empirical questions and that when the results from these procedures are in agreement, greater confidence can be placed in the internal structure of a test.

As in other investigations (e.g., Canivez, 2013a; McGill & Busse, 2015), multicollinearity of the FCI and the factor scores in the hierarchical multiple regression analyses

was observed in the present study due to the linear combination of subtests to produce factor scores and the FCI. However, this redundancy is precisely the problem that practitioners must confront when simultaneously interpreting full-scale and factor-level scores on intelligence tests such as the KABC-II. Additionally, it should be noted that multicollinearity is not a threat to validity in regression studies that are limited to interpreting the R^2 statistic (Cohen, Cohen, West, & Aiken, 2003), nor does it invalidate the use of hierarchical multiple regression analysis to detect improvements in R^2 such as those provided by the CHC factor scores beyond the FCI (Schneider, 2008).

Conclusion and Implications for Clinical Practice

As noted by Schneider (2013), it is important for clinicians to understand that group factor scores on IQ tests measure more discrete cognitive tasks at a lower level of generality and thus they are imperfectly correlated. As a result, unity within a cognitive profile should not be expected. However by focusing interpretation on a lower level of dimensionality (e.g., Stratum II broad abilities), clinicians disregard the effects of the higher-order general factor that accounts for a significant amount of the variation in cognitive test scores (Beaujean, 2015; Beaujean, Parker, & Parkin, 2014; Gignac, 2016) as well as the appropriate structuring of cognitive variables (Carroll, 1993; Schneider & McGrew, 2012). The present results show that invalidation of the global composite on the KABC-II in the presence of significant scatter does not obviate its influence on lower-order factor and subtest scores and thus should not automatically be disregarded in those circumstances to comport with a popular *rule of thumb*. As noted by Nickerson (2004), such decision-making heuristics "can become bothersome when their limitations are not borne in mind" (p. 371).

Despite the nascent empirical support for the variability hypothesis and the popular interpretive recommendations that stem from it (e.g., Groth-Marnat, 2009; Hale & Fiorello, 2004), it continues to be endorsed by assessment specialists and widely applied in training and practice on the basis of its intuitive appeal and the supposition that significant scatter is rare and thus a useful focal point for clinical conjecture (Gambrill, 2012; Watkins, 2000; 2003). However, it has long been noted that impressionistic and subjective judgement are not an appropriate substitute for empirical evidence as a source of knowledge regarding the efficacy of clinical practices (Lilienfeld, Ammirati, & David, 2012; McFall, 2000). While additional research is needed to more fully examine the effects of scatter on related measurement instruments and the potential implications for clinical interpretation of IQ scores, clinicians are advised to reconsider the call to automatically invalidate the FSIQ when scatter is observed as a principle that should *never* be violated until additional empirical evidence is provided to support this practice (Baron, 2008). That is not to suggest that the FSIQ should be the sole focus of interpretation in clinical assessment as evidence was found to support interpretation beyond *g* in certain circumstances. However, additional interpretation of group factor scores should be undertaken only after careful considering their precision of measurement and potential clinical utility (DeMars, 2013). When making clinical decisions with these data, psychologists are encouraged to consider Weiner's advice to assessors to "(a) know what their tests can do and (b) act accordingly" (1989, p. 829).

About the Author

Ryan J. McGill, Ph.D., BCBA-D, NCSP is an assistant professor in the school psychology program at The College of William and Mary. Correspondence concerning this article should be addressed to Ryan J. McGill, School of Education, The College of William and Mary, P. O. Box 8795, Williamsburg,Virginia 23187. **E-mail:** *rmcgill@wm.edu*

References

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

Bain, K., & Gray, R. (2008). Test reviews: Kaufman, A. S., & Kaufman, N. L. (2004). Kaufman Assessment Battery for Children, Second edition. Circle Pines, MN: AGS. *Journal of Psychoeducational Assessment, 26,* 92-101. doi: 10.1177/0734282907300461

Baron, J. (2008). *Thinking and deciding* (4th ed.). New York: Cambridge University Press.

Beall, A. L., Holdnack, J. A., Saklofske, D. H., & Weiss, L. G. (2016). Practical considerations in WISC-V interpretation and intervention. In L. G. Weiss, D. H. Saklofske, J. A. Holdnack, & A. Prifitera (Eds.), *WISC-V assessment and interpretation: Scientist-practitioner perspectives*. (pp. 66-93). San Diego, CA: Academic Press.

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., Parkin, J., & Parker, S. (2014). Comparing Cattell-Horn-Carroll factor models: Differences between bifactor and higher order factor models in predicting language achievement. *Psychological Assessment, 26,* 789-805. doi: 10.1037/a0036745

Braden, J. P. & Ouzts, S. M. (2005). Review of Kaufman Assessment Battery for Children, Second Edition. In R. A. Spies & B. S. Plake (Eds.), *The sixteenth mental measurements yearbook* (pp. 517-520). Lincoln: University of Nebraska.

Canivez, G. L. (2008). Orthogonal higher-order factor structure of the Stanford-Binet Intelligence Scales for children and adolescents. *School Psychology Quarterly, 23,* 533-541. doi: 10.1037/a0012884

Canivez, G. L. (2011). Hierarchical factor structure of the Cognitive Assessment System: Variance partitions from the Schmid-Leiman (1957) procedure. *School Psychology Quarterly, 26,* 305-317. doi: 10.1037/a0025973

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. Schwean (Eds.). *The Oxford handbook of child psychological assessment* (pp. 84-112). New York: Oxford University Press.

Canivez, G. L. (2014). Construct validity of the WISC-IV with a referred sample: Direct versus indirect hierarchical structures. *School Psychology Quarterly, 29,* 38-51. doi: 10.1037/spq0000032

Canivez, G. L., & Watkins, M. W. (2010). Investigation of the factor structure of the Wechsler Adult Intelligence Scale - Fourth Edition (WAIS-IV): Exploratory and higher-order factor analyses. *Psychological Assessment, 22,* 827-836. doi: 10.1037/a0020429

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: 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. (1998). Human cognitive abilities: A critique. In J. J. McArdle & R. W. Woodcock (Eds.), *Human cognitive abilities in theory and practice* (pp. 5-23). 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. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245-276. doi: 10.1207/s15327906mbr0102_10

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72,* 193-204. doi: 10.1037/h0024670

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J. P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality, 80,* 219-251. doi:10.1111/j.1467–6494.2011.00739.x

Child, D. (2006). *The essentials of factor analysis* (3rd ed.). New York: Continuum Publishing.

Cohen, J. (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). New York: Psychology Press.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum

Courville, T., Coalson, D. L., Kaufman, A. S., & Raiford, S. E. (2016). Does WISC-V scatter matter? In A. S. Kaufman, S. E. Raiford, & D. L. Coalson, (Eds.), *Intelligent testing with the WISC-V* (pp. 209-228). Hoboken, NJ: John Wiley.

Dana, J., & Dawes, R. (2007). Comment on Fiorello et al. Interpreting Intelligence Test Results for Children with Disabilities: Is Global Intelligence Relevant? *Applied Neuropsychology, 14,* 21-25. doi: 10.1080/09084280701280379

Daniel, M. H. (2007). "Scatter" and the construct validity of FSIQ: Comment on Fiorello et al. (2007). *Applied Neuropsychology, 14,* 291-295. doi: 10.1080/09084280701719401

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgement. *Science, 243,* 1668-1674. doi: 10.1126/science.2648573

Decker, S. L., Hale, J. B., & Flanagan, D. P. (2013). Professional practice issues in the assessment of cognitive functioning for educational applications. *Psychology in the Schools, 50,* 300-313. doi: 10.1002/pits.21675

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing, 13,* 354-378. doi: 10.1080/15305058.2013.799067

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., & Gischlar, K. L. (2014). Ethical and empirical considerations in the identification of learning disabilities. *Journal of Applied School Psychology, 30,* 68-82. doi: 10.1080/15377903.2013.869786

Dombrowski, S. C., Canivez, G. L., Watkins, M. W., & Beaujean, A. A. (2015). Exploratory bifactor analysis of the Wechsler Intelligence Scale for Children-Fifth Edition with the 16 primary and secondary subtests. *Intelligence, 53,* 194-201. doi: 10.1016/j.intell.2015.10.009

Drozdick, L. W., Wahlstrom, D., Zhu, J., & Weiss, L. G. (2012). The Wechsler Adult Intelligence Scale-Fourth Edition and the Wechsler Memory Scale-Fourth Edition. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (3rd ed.; pp. 197-223). New York: Guilford Press.

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. New York: Oxford University Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4,* 272-299. doi: 10.1037/1082-989X.4.3.272

Faust, D. (1990). Data integration in legal evaluations: Can clinicians deliver on their premises? *Behavioral Sciences and the Law, 7,* 469-483. doi: 10.1002/bsl.2370070405

Feifer, S. G., Nader, R. G., Flanagan, D. P., Fitzer, K. R., & Hicks, K. (2014). Identifying specific reading subtypes for effective educational remediation. *Learning Disabilities: A Multidisciplinary Journal, 20,* 18-30. Retrieved from http://js.sagamorepub.com/ldmj/article/view/5151

Fiorello, C. A., Hale, J. B., Holdnack, J. A., Kavanagh, J. A., Terrell, J., & Long, L. (2007). Interpreting intelligence test results for children with disabilities: Is global intelligence relevant? *Applied Neuropsychology, 21*. 2-12. doi: 10.1080/09084280701280338

Fiorello, C. A., Hale, J. B., McGrath, M., Ryan, K., & Quinn, S. (2002). IQ interpretation for children with flat and variable test profiles. *Learning and Individual Differences, 13,* 115-125. doi: 10.1016/S1041-6080(02)00075-4

Fiorello, C. A., Hale, J. B., & Wycoff, K. L. (2012). Cognitive hypothesis testing. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 484-496). New York: Guilford.

Flanagan, D. P., & Kaufman, A. S. (2004). *Essentials of WISC-IV assessment*. Hoboken, NJ: John Wiley.

Flanagan, D. P., Alfonso, V. C., & Mascolo, J. T. (2011). A CHC-based operational definition of SLD: Integrating multiple data sources and multiple data-gathering methods. In D. P. Flanagan & V. C. Alfonso (Eds.), *Essentials of specific learning disability identification* (pp. 233-298). Hoboken, NJ: John Wiley.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). *Essentials of cross-battery assessment* (3rd ed.). Hoboken, NJ: John Wiley.

Flanagan, D. P., Ortiz, S. O., Alfonso, V. C., & Dynda, A. (2006). Integration of response-to-intervention and norm-referenced tests in learning disability identification: Learning from the Tower of Babel. *Psychology in the Schools, 43,* 807-825. doi: 10.1002/pits.20190

Fletcher, J. M., Steubing, K. K., Morris, R. D., & Lyon, G. R. (2013). Classification and definition of learning disabilities: A hybrid model. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (2nd ed., pp. 33-50). New York: Guilford Press.

Frazier, T. W., & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? *Intelligence, 35,* 169-182. doi: 10.1016/j.intell.2006.07.002

Freberg, M., Vandiver, B. J., Watkins, M. W., & Canivez, G. L. (2008). Significant factor score variability and the validity of the WISC-III full scale IQ in predicting later academic achievement. *Applied Neuropsychology, 15,* 131-139. doi: 10.1080/09084280802084010

Gambrill, E. (2012). *Critical thinking in clinical practice: Improving the quality of judgements and decisions* (3rd ed.). Hoboken, NJ: John Wiley.

Gignac, G. E. (2007). Multi-factor modeling in individual differences research: Some recommendations and suggestions. *Personality and Individual Differences, 42,* 37-48. doi: 0.1016/j.paid.2006.06.019

Gignac, G. E. (2016). Residual group-level factor associations: Possibly negative implications for the mutualism theory of general intelligence. *Intelligence, 55,* 69-78. doi: 10.1016/j.intell.2016.01.007

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., McDermott, P. A., Watkins, M. W., & Kush, J. C. (1997). The base rate problem and its consequences for interpreting children's ability profiles. *School Psychology Review, 26,* 176-188. Retrieved from http://www.nasponline.ord

Glutting, J. J., Watkins, M. W., & Youngstrom, E. A. (2003). Multifactored 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: Guilford.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Gorsuch, R. L. (2003). Factor analysis. In J. A. Schinka & F. F. Velicer (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 143-164). Hoboken, NJ: John Wiley.

Gottfredson, L. S. (1997). Why *g* matters: The complexity of everyday life. *Intelligence, 24,* 79-132. doi: 10.1016/S0160-2896(97)90014-3

Groth-Marat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: John Wiley.

Hale. J. B., & Fiorello. C. A. (2001). Beyond the academic rhetoric of 'g': Intelligence testing guidelines for practitioners. *The School Psychologist, 55,* 113-139. Retrieved from http://www.apadivisions.org/division-16/publications/newsletters/school-psychologist/2001/11-issue.pdf

Hale, J. B., & Fiorello, C. A. (2004). *School neuropsychology: A practitioner's handbook*. New York: Guilford.

Hale, J. B., Fiorello, C. A., Kavanuagh, J. A., Holdnack, J. A., & Aloe, A. M. (2007). Is the demise of IQ interpretation justified? A response to special issue authors. *Applied Neuropsychology, 14,* 37-51. doi:10.1080/09084280701280445.

Hale, J. B., Fiorello, C. A., Dumont, R., Willis, J. O., Rackley, C., & Elliott, C. (2008). Differential Ability Scales-Second Edition (neuro)psychological predictors of math performance for typical children and children with math disabilities. *Psychology in the Schools, 45,* 838-858. doi: 10.1002/pits.20330

Haynes, S. N., Smith, G. T., & Hunsley, J. D. (2011). *Scientific foundations of clinical assessment*. New York: Routledge.

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

Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. McGrew, J. K. Werder, & R. W. Woodcock, *WJ-R technical manual* (pp. 197–232). Itasca, IL: Riverside Publishing.

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

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39,* 31-36. doi: 10.1007/BF02291575

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III*. New York: John Wiley.

Kaufman, A. S., & Kaufman, N. L. (2004a). *Kaufman Assessment Battery for Children* (2nd ed.). Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (2004b). *Kaufman Assessment Battery for Children-Second Edition manual*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (2004c). *Kaufman Test of Educational Achievement* (2nd ed.). Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (2004d). *Kaufman Test of Educational Achievement-Second Edition manual*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., Lichtenberger, E. O., Fletcher-Janzen, E., & Kaufman, N. L. (2005). *Essentials of KABC-II assessment*. Hoboken, NJ: John Wiley.

Kotz, K. M., Watkins, M. W., & McDermott, P. A. (2008). Validity of the general Conceptual Ability score on the Differential Ability Scales as a function of significant and rare

interfactor variability. *School Psychology Review, 37,* 261-278. Retrieved from http://www.nasponline.org

Kranzler, J. H., & Floyd, R. G. (2013). *Assessing intelligence in children and adolescents: A practical guide*. New York: Guilford Press.

Lichtenberger, E. O., Sotelo-Dynega, M., & Kaufman, A. S. (2009). The Kaufman Assessment Battery for Children-Second Edition. In J. A. Naglieri & S. Goldstein (Eds.), *Practitioner's guide to assessing intelligence and achievement* (pp. 61-94). Hoboken, NJ: John Wiley.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). New York: Oxford University Press.

Lilienfeld, S. O., Ammirati, R., & David, M. (2012). Distinguishing science from pseudoscience in school psychology: Science and scientific thinking as safeguards
against human error. *Journal of School Psychology, 50,* 7-36. doi:
10.1016/j.jsp.2011.09.006
Luria, A. R. (1966). *Human brain and psychological processes*. New York: Harper Row.

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

Marley, S. C., & Levin, J. R. (2011). When are prescriptive statements in educational research justified? *Educational Psychology Review, 23*, 197-206. doi: 10.1007/s10648-011-9154-y

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. *The Journal of Special Education, 25,* 504-526. doi: 10.1177/002246699202500407

McDermott, P. A., Watkins, M. W., & Rhoad, A. (2014). Whose IQ is it?—Assessor bias variance in high-stakes psychological assessments. *Psychological Assessment, 26,* 207-214. Doi: 10.1037/a0034832

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

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). Interpretation of KABC-II scores: An evaluation of the incremental validity of Cattell-Horn-Carroll (CHC) factor scores in predicting achievement. *Psychological Assessment, 27,* 1417-1426. doi: 10.1037/pas0000127

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., & Spurgin, A. R. (2015). Exploratory higher order analysis of the Luria interpretive model on the Kaufman Assessment Battery for Children-Second Edition (KABC-II) school-age battery. *Assessment*. Advanced online publication. doi: 10.1177/1073191115614081

McGill, R. J., Styck, K. M., Palomares, R. S., & Hass, M. R. (2015). Critical issues in specific learning disability identification: What we need to know about the PSW model. *Learning Disability Quarterly*. Advance online publication. doi: 10.1177/0731948715618504

Meehl, P .E. (1956). Wanted - a good cook-book. *American Psychologist, 11,* 263-272. doi: 10.1037/h0044164

Naglieri, J. A. (2000). Can profile analysis of ability test scores work? An illustration using the PASS theory and CAS with an unselected cohort. *School Psychology Quarterly, 15,* 419-433. doi: 10.1037/h0088798

Nelson, J. M., & Canivez, G. L. (2012). Examination of the structural, convergent, and incremental validity of the Reynolds Intellectual Assessment Scales (RIAS) with a clinical sample. *Psychological Assessment, 24,* 129-140. doi:10.1037/a0024878

Nickerson, R. S. (2004). *Cognition and Chance: The psychology of probabilistic reasoning*. Mahwah, NJ: Erlbaum.

Pfeiffer, S. I., Reddy, L. A., Kletzel, J. E., Schmelzer, E. R., & Boyer, L. M. (2000). The practitioner's view of IQ testing and profile analysis. *School Psychology Quarterly, 15,* 376-385. doi: 10.1037/h0088795

Prifitera, A., & Dersh, J. (1993). Base rates of WISC-III diagnostic subtest patterns among normal, learning-disabled, and ADHD samples. *Journal of Psychoeducational Assessment Monograph Series, 43-55.*

Rapaport, D., Gil, M., & Schafer, R. (1945). *Diagnostic psychological testing: The theory, statistical evaluation, and diagnostic application of a battery of tests* (Vol. 1). Chicago: Yearbook Medical.

R Development Core Team. (2016). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

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

Reschly, D. J., Myers, T. J., & Hartel, C. (Eds.). (2002). *Mental retardation: Determining eligibility for social security benefits*. Washington, DC: National Academy Press.

Revelle, W. (2012). *Psych: Procedures for psychological, psychometric, and personality research (version 1.2.4)* [Computer Software]. Evanston, IL: Northwestern University

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

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86,* 162-173. doi: 10.1037/0022-3514.86.1.162

Schneider, W. J. (2008). Playing statistical Ouija board with communality analysis: Good questions, wrong assumptions. *Applied Neuropsychology, 15,* 44-53. doi: 10.1080/09084280801917566

Schneider, W. J. (2013). Principles of assessment aptitude and achievement. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwean (Eds.). *The Oxford handbook of child psychological assessment* (pp. 286-330). New York: Oxford University Press.

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.

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

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Association.

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

Wasserman, J. D., & Bracken, B. A. (2013). Fundamental psychometric considerations in assessment. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Assessment psychology* (2nd ed., Vol. 10, pp. 50-81). Hoboken, NJ: John Wiley.

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. Retrieved from http://www.srmhp.org/0202/iq.html

Watkins, M. W. (2004). *MacOrtho.* [Computer Software]. State College, PA: Ed & Psych Associates.

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. (2013). *Omega*. [Computer Software]. Phoenix, AZ: Ed & Psych Associates.

Watkins, M. W., Glutting, J. J., & Lei, P. W. (2007). Validity of the full scale IQ when there is significant variability among WISC-III and WISC-IV factor scores. *Applied Neuropsychology, 14,* 13-20. doi: 10.1080/09084280701280353

Watkins, M. W., & Smith, L. G. (2013). Long-term stability of the Wechsler Intelligence Scale for Children-Fourth Edition. *Psychological Assessment, 25,* 477-483. doi: 10.1037/a0031653

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale - Fourth Edition technical and interpretive manual*. San Antonio, TX: NCS Pearson.

Wechsler, D. (2014). *Wechsler Intelligence Scale for Children - Fifth Edition technical and interpretive manual*. San Antonio, TX: NCS Pearson.

Weiner, I. B. (1989). On competence and ethicality in psychodiagnostic assessment. *Journal of Personality Assessment, 53,* 827-831. doi:10.1207/s15327752jpa5304_18

Wiggins, J. S. (1988). *Personality and prediction: Principles of personality assessment*. Malabar, FL: Krieger Publishing.

Wood, J. M., Tatryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods, 1,* 354-365. doi: 10.1037/1082-989X.1.4.354

Yates, B. T., & Taub, J. (2003). Assessing the costs, benefits, cost-effectiveness, and cost-benefit of psychological assessment: We should, we can, and here's how. *Psychological Assessment, 15,* 478-495. doi: 10.1037/1040-3590.15.4.478