Chapter 3

Kaufman Assessment Battery for Children–Second Edition/Normative Update (KABC–II/ KABC–II NU): Clinical Interpretation From an Evidence-Based Perspective[1]

Ryan J. McGill and Stefan C. Dombrowski

The *Kaufman Assessment Battery for Children–Second Edition* (KABC–II; Kaufman & Kaufman, 2004a) measures the intellectual and processing abilities of children and adolescents ages 3:0 to 18:11 years and contains a total of 18 subtests.[2] The KABC–II was a major revision of the original K–ABC (Kaufman & Kaufman, 1983), which was a revolutionary addition to the suite of cognitive testing instrumentation available to practitioners at that time. The original K–ABC was designed principally from the standpoint of neuropsychological theory; in particular, Luria's depiction of cognitive processing (Luria, 1966). This model posits that the brain is composed of three functional units. The first unit is made up largely of older areas of the brain (i.e., the brain stem and reticular activation system) and is concerned with promoting alertness to external stimuli; the second unit is formed by the parietal, occipital and temporal lobes and is responsible for storing and processing sensory information from the environment; and the third unit is formed by the frontal lobe, which is responsible for executive functions and planning behavior in response to external stimuli. For the K–ABC, Luria's model was repurposed in the form of a sequential-versus-simultaneous processing framework, and when combined with additional measures of academic functioning, it represented a significant departure from the traditional ability testing zeitgeist that had existed to that time (Keith, 1985). Additionally, the K–ABC was unique in that it established a new benchmark for the reporting of psychometric information in its technical manual and that several graduate students, who would go on to be

considered luminaries in assessment psychology in their own right, participated in its development under Alan Kaufman's supervision at the University of Georgia (Kaufman, 2009).

Despite these achievements, questions were soon raised about the veracity of the instrument's theoretical structure. Most notably, Keith (1985) conducted exploratory and confirmatory analysis on the normative sample and produced results suggesting that the cognitive processing elements posited by the test publisher (i.e., sequential-simultaneous processing) was largely supported across age groups. However, it was suggested that those dimensions could potentially be alternatively conceptualized as Verbal Memory and Nonverbal Reasoning. More concerning, the measures of achievement did not cohere with the identified cognitive dimensions, thus calling into question the nature of the general factor thought to be measured by the omnibus full-scale score. Bracken (1985) further criticized the lack of alignment between the test and its presumed theoretical underpinnings noting, "Faults may be found with the instrument, but commendably, most of the necessary data is made available in the K–ABC manual that enables one to find the faults" (p. 35).

The K–ABC was eventually revised with the publication of the KABC–II in 2004. The KABC–II underwent a major structural revision with eight subtests being eliminated and 10 indicators created and added to the new battery. The KABC–II is unique in that it is the only commercial ability measure to utilize a dual-interpretive structure: the Cattell-Horn-Carroll theory of human cognitive abilities (CHC; Schneider & McGrew, 2018) and Luria's theory of cognitive processing (Luria, 1966). This affords flexibility to the examiner in which approach to utilize for any examinee though the technical manual (Kaufman & Kaufman, 2004b) stresses that an interpretive preference must be selected *a priori* so as to prevent an examiner from morphing from one battery to another to suit their preferences. Before elaborating on the differences

between the CHC and Luria models, it is important to point out that the original K–ABC also employed a hybrid interpretive approach in which elements of Luria's theory were embedded within a more global Fluid-Crystallized, or Gf-GC theory (Horn & Cattell, 1966) context, a linkage that does not appear to have been recognized in the literature concerning the instrument.

The CHC model features 10 core subtests that contribute to the measurement of five group-specific factors (Crystallized Ability [Gc], Fluid Reasoning [Gf], Visual Processing [Gv], Long-Term Storage and Retrieval [Glr], and Short-Term Memory [Gsm])[3] and an omnibus, full-scale score termed the *Fluid-Crystallized Composite* (FCI). By contrast, the Luria model employs a more parsimonious eight-core subtest configuration contributing to the measurement of four group-specific factors (Planning, Sequential Processing, Simultaneous Processing, and Learning) and a full-scale score termed the *Mental Processing Index* (MPI). An optional Delayed Recall index can also be calculated but it not structurally derived; therefore, its veracity is questionable. The hypothesized measurement models for these dueling interpretive structures at school age are depicted in Figures 3.1 and 3.2 respectively. As noted by McGill (2017), the only salient difference between the models is that the Luria model omits measures of acquired knowledge, also known as *Crystallized Ability* within the CHC nomenclature. It should also be noted that Fluid Reasoning was not able to be consistently located at preschool age, thus a more parsimonious scoring structure is employed across that age range. Additional supplementary subtests whose configurations differ according to the age of the child can also be administered but they do not contribute to the measurement of core global scales or indices.

<F03_001>

<F03_002>

Although users are advised in the manual (Kaufman & Kaufman, 2004b) to prioritize interpreting the test from the CHC perspective, the Luria model may be preferred in specific clinical situations including, but not limited to, examining children from culturally and linguistically diverse backgrounds or assessing individuals with known language impairments (Drozdick et al., 2018). It should also be noted that the KABC–II also provides users with an alternative nonverbal index of overall ability that is based on different configurations of core and supplementary measures that omit verbal responding from the examinee. As a result of this versatility, it is not surprising that surveys of the contemporary assessment practice of school psychologists consistently reveal that the KABC–II is one of the most often used instruments for the clinical assessment of intelligence and particularly among examiners assessing the cognitive abilities of culturally and linguistically diverse minorities (Benson et al., 2019; Sotelo-Dynega & Dixon, 2014).

In the remainder of this chapter, the suggested KABC–II interpretive procedures will be outlined followed by a review of the validation evidence presented in the manual to support the use of those procedures. We will then conclude with a review of selected independent examinations of various aspects of construct validity for the instrument and its scores. For the sake of parsimony, particular attention will be paid to structural fidelity, predictive validity, and diagnostic utility (Keith & Kranzler, 1999). Our focus is limited to the version of the instrument normed for use in the United States.[4]

Suggested Interpretive Procedures for the Instrument

Like other commercial ability measures, a detailed step-by-step series of the interpretive procedures is outlined in the *KABC–II Technical Manual* (Kaufman & Kaufman, 2004b). As previously noted, users should have already selected the preferred interpretive model prior to

administering the instrument to the examinee. Once those scores are obtained, users are encouraged to start by interpreting the omnibus full-scale score[5] before evaluating for index level strengths and weaknesses. However, "whether the FCI or MPI is used, before evaluating the global score you need to determine whether the global score is interpretable" (Kaufman et al., 2005). In doing so, examiners are invited to employ an intuitive rule of thumb that if the difference between the highest and lowest index score meets or exceeds 23 points then the global score should not be interpreted. Interpretability is also a consideration at the index level, and Kaufman and colleagues (2005) suggest that interpretation of an index score should only occur if "the child performed consistently on the Core subtests that compose that scale" (p. 89). There is no empirical evidence to support the use of these scatter heuristics in the technical manual or in the accompanying KABC–II interpretive literature.

While not required, examiners also have the option of conducting one or more of the plethora of planned clinical comparisons between index- and subtest-level scores that have been outlined in KABC–II interpretive resources (e.g., Drozdick et al., 2018; Kaufman et al., 2005), for example, combining different subtests to create pseudo-composites that compare a child's presumed performance across meaningful versus abstract stimuli. Finally, the KABC–II asks users to note relevant test session behaviors in the form of a qualitative checklist at the end of each subtest. These behaviors can be tabulated in the test record form, and users are invited to consider the potential clinical relevance of these tabulations.

Issues with the Development and Validation of the KABC–II

The *KABC–II Technical Manual* (Kaufman & Kaufman, 2004b) reports a wealth of reliability and validity evidence and provides a detailed description of the theoretical development of the instrument. Despite these strengths, several psychometric shortcomings and

questions were noted in a recent factor analytic study by McGill (2020). Kaufman and Kaufman reported, “The KABC–II development process relied mostly on the technique of confirmatory factor analysis, used in an *exploratory* fashion[6] to evaluate subtests and decide how they should be grouped into scales” (p. 103). This is a curious disclosure given the fact that the theoretical rationale for the KABC–II appears to be well-established in the psychometric literature that was available at that time.

A major focus of the factor analyses conducted on the normative sample was to determine if Fluid Reasoning indicators could be separated from the Visual Processing measures. Whereas it is reported that EFA was used to identify possible alternative interpretive structures for the test, the EFA results “did not make a significant contribution to the overall analysis program” (p. 104) and were not reported in the technical manual. Instead, the test publisher relied exclusively on confirmatory factor analysis (CFA) employed across two stages. In stage one, CFA was used to examine various higher-order models ranging from a one-factor model to increasingly more complex correlated (oblique) factor models at each age range. The authors report checking at each step to verify that improvement in fit was statistically significant and that there was no evidence of local model misspecification (i.e., problematic loadings, out of bounds parameter estimates). Fit statistics for the models that were explored at this stage are not reported but the results for each age are described narratively. Nevertheless, it appears that there were issues getting all of the hypothesized CHC dimensions to emerge consistently across the age range if at all.

At age three, a one-factor (*g*) model was preferred even though it is reported that a Short-Term Memory dimension could be located that appeared to meet the stated inclusionary criteria in the manual. At age four, analyses supported the presence of Short-Term Memory and Long-

Term Storage and Retrieval, but a Fluid Reasoning factor could not be located. As a result, hypothesized Fluid Reasoning measures were alternatively assigned to the Visual Processing factor. Although it can certainly be argued that it is likely that Fluid Reasoning could not be located at that age because of developmental differences in preschool children, it is not clear how the measures supposedly contributing to the measurement of Visual Processing morph to another dimension later in the developmental span. Interestingly, it was noted that the Crystallized Ability and Visual Processing factors were highly correlated, suggesting that they were likely isomorphic. Even so, a decision was made to retain both dimensions even though it is acknowledged that they were not statistically significant. Instead, the decision to retain them was based on consideration of their qualitative content. Nevertheless, the dimensional complexity of these measures raises the question of the discriminant validity of several of the hypothesized CHC dimensions.

At ages 5–7, the Fluid Reasoning and Visual Processing factors were not distinguishable, and all those measures were again assigned to the same factor. It was later reported that the latent factor correlations between Visual Processing and Fluid Reasoning exceeded .90 across the school age span (7–18), raising concern about whether Fluid Reasoning is a viable dimension for the instrument (Byrne, 2005). However, the authors noted that these abilities could be differentiated statistically, and the decision was made to retain Fluid Reasoning to better cohere with the presumed CHC interpretive model for the test. Additionally, different subtest configurations for the Visual Processing scale were explored to determine the most optimal combination of core tests that would help distinguish the abilities from each other. Results indicated that Triangles and Rover provided the most optimal differentiation in the younger

portion of the school age range (7–12), and Block Counting and Rover was the best combination at ages 13 to 18.

In the next stage of the validation plan, a series of constrained CFAs were used to investigate hierarchical versions of the models retained at stage one containing a second-order general factor of intelligence. Those models are depicted visually in Figures 8.1 and 8.2 of the technical manual (pp. 106–107). It was believed that these models best aligned with the CHC interpretive structure for the test. Inspection of global fit statistics indicate that all the models fit the data well. However, at ages 7–18, all the models contain path loadings between *g* and Fluid Reasoning that suggest empirical redundancy.

As noted by Brown (2015), whereas standardized path loadings equal to 1.0 are technically permissible in CFA, they present an interpretive confound as they indicate that the lower-order dimension does not account for any meaningful variance that distinguishes it from general intelligence. When these problematic loadings are encountered, researchers are frequently advised in the methodological literature to delete the redundant variable(s) in accordance with the scientific law of parsimony. Values that exceed 1.0 are considered out-of-bounds estimates (i.e., Heywood cases) and indicate potential model misspecification. In sum, the factor analytic evidence reported in the technical manual (Kaufman & Kaufman, 2004b) raises concern about the veracity of the preferred CHC interpretive model. Additionally concerning are the correlations with similar CHC measures from the *Woodcock-Johnson IV Tests of Cognitive Abilities* (WJ IV) reported in that test's technical manual (McGrew et al., 2014). For example, the reported correlations between like Fluid Reasoning and Visual Processing scores was .46 and .37, respectively, which is relatively low. By contrast the correlation between like measures of Crystallized Ability was .82. If these indices from the KABC–II are *actually*

measuring these abilities, then one would expect the alignment between what are purported to be estimates of the same latent dimension to be much stronger. Finally, the structural integrity of the Luria interpretive model was not explored in any meaningful way, which is problematic because it cannot be automatically assumed that a previously established higher-order solution will be maintained when the configuration of observed variables is altered.

Variance Partitioning and the Interpretive Relevance of Lower-Order Dimensions

Even if consensus is achieved with respect to ascertaining the correct number of first-order dimensions measured by the KABC–II, additional information is necessary for determining the interpretive relevance of the scores aligned with those dimensions. It is important to keep in mind that in a hierarchical measurement model, first-order factors are abstractions of measured variables; thus, extrapolating a higher-order general factor represents an abstraction from an abstraction (Beaujean, 2015). More importantly, a non-trivial portion (often the vast majority) of reliable variance in all subtests and first-order dimensions is attributable to general intelligence. Failing to consider this source of influence when interpreting first-order factors will lead to overestimating the effects of those attributes in explaining performance on the KABC–II (Caretta & Ree, 2001). As a result, Carroll (1995) insisted that it is necessary to decompose variance into components that can be sourced more appropriately to higher- and lower-order dimensions. To accomplish this task, he recommended second-order factor analysis of first-order factor correlations followed by a Schmid-Leiman (SL; Schmid & Leiman, 1957) procedure. When applied to factor analytic solutions, the SL procedure allows for the calculation of first-order subtest loadings that are independent of the influence of a higher-order general factor. 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).

More recently, methodologists have encouraged the use of bifactor modeling in CFA (Reise, 2012) to examine these effects. Although it has been argued that the SL procedure represents an approximate bifactor model in EFA, it is merely a reparameterization of the hierarchical model (Canivez, 2016). However, over the last decade, the two techniques have been used interchangeably in the psychometric literature to partition variance in cognitive tests (Dombrowski, McGill, Canivez et al., 2021).

Whether produced from a pure bifactor CFA model or SL procedure in EFA, orthogonalized factor loading estimates can also be used to produce various indices that evaluate dimensionality and aid in the evaluation of whether a particular score is clinically relevant. Although each of these estimates is important in its own right, Omega coefficients are often the focal point in determining whether a factor can be interpreted with confidence in clinical practice. Omega-hierarchical (ω_H) and omega-hierarchical subscale (ω_{HS}) estimate the unit-weighted portion of reliable variance in latent factors. The ω_H coefficient is the estimate for the general intelligence factor with variability of group factors removed, while the ω_{HS} coefficient is the estimate of a group factor with all other group and the general factor removed (Rodriguez et al., 2016). Although subjective, it has been suggested that omega coefficients should at a minimum exceed .50, but .75 is preferred (Reise et al., 2013). Additionally, it is important to

consider explained common variance (ECV) and construct replicability (*H*). If the KABC–II factor or subtest scores fail to capture meaningful portions of true score variance they will likely be of limited clinical utility. Unfortunately, this information is not reported in the technical manual.

Post-Publication KABC–II Psychometric Evidence

Given these limitations, the KABC–II has been the subject of numerous psychometric investigations. In particular, investigations that have sought to clarify its interpretive structure. Next we conduct a selected review of major KABC–II construct validity studies. Consistent with recent calls for advancing the cause of evidence-based practice to the practice of clinical assessment, we approach this evidence from an *evidence-based assessment* perspective (EBA; Dombrowski, 2020; Youngstrom, 2013).

Evidence-Based Assessment (EBA)

Practitioners and scholars have access to a plethora of information that may be useful for informing clinical assessment practice. Unfortunately, it is difficult to determine the specific sources of information that are most useful for informing the clinical bottom line. According to Hunsley and Mash (2007), "a truly evidence-based approach to assessment, therefore, would involve an evaluation of the accuracy and usefulness of this complex decision-making task" (p. 30) when considering the high degree of error endemic within this process. The EBA framework goes beyond traditional psychometric reliability and validity evidence and focuses more on the outcomes and utility of assessment processes. According to Youngstrom (2013), traditional approaches to scale validation remain important, but what really matters are the 3 *P*s of clinical assessment: prediction (assessment data's relationship to important external criteria); prescription (treatment utility of assessment); and process (identification of mediating variables

for treatment). Essentially the goal of EBA is to subject our assessment method and clinical decision-making models to risky empirical tests (Meehl, 1978) to separate the proverbial wheat from the chaff in assessment science.

CHC Model Structural Validity

Most independent examinations of the latent structure of the KABC–II have focused on validating the preferred CHC interpretive model. For example. Reynolds and colleagues (2007) examined the measurement invariance of the KABC–II theoretical structure to ascertain the degree to which it was consistent across age and to verify the instrument's consistency with CHC theory. Results indicated the measurement model was invariant across age after imposing constraints to account for the fact that several subtests cannot be administered at certain ages. However, the model that was explicated is not described. A series of rival CHC models were then sequentially explored. Consistent with the CFA results reported in the technical manual, the baseline model contained evidence of a Heywood case and is not tenable for the data. The final validation model departed from publisher theory, most notably reporting a cross-loading for Pattern Reasoning on Visual Processing and Fluid Reasoning.

Later the KABC–II was included as one of five test batteries in a cross-battery (XBA; Flanagan et al., 2013) reference variable investigation of the CHC taxonomy featuring 423 participants ages 6 to 16 years (Reynolds et al., 2013). Results generally supported a three-stratum factor relatively consistent with a priori theory. However, despite attempts to apply post hoc constraints to the model, Fluid Reasoning could not be statistically distinguished from *g* though it was featured in the model figures that were reported and not eschewed as would be consistent with best practice (Brown, 2015). Additionally, the Pattern Reasoning cross-loading

identified by Reynolds et al. (2007) was not modeled. The results were replicated in a more recent XBA study by Caemmerer et al. (2020).

Inexplicably, took nearly 14 years before the KABC–II was subjected to EFA. In their analyses, McGill and Dombrowski (2018) examined both the core and supplementary subtest structures at school age (7–18) using the data obtained from normative participants. Core battery results supported a four-factor solution unifying the Fluid Reasoning and Visual Processing indicators and featuring a more accurately named Perceptual Reasoning factor. Inclusion of the supplementary measures provided support for a five-factor model that departed significantly from publisher theory. Whereas the Perceptual Reasoning factor was again supported, the addition of a fifth factor resulted in the theoretically preferred Long-Term Storage and Retrieval factor splitting into two different Glr dimensions. More concerning, ω_H coefficients for the general factor all exceeded .80 indicating that dimension possesses enough reliable variance to be interpreted. However, none of the ω_{HS} estimates for the group-specific factors exceeded .47 indicating that those dimensions cannot be interpreted with confidence.

In 2018 the KABC–II was subjected to a normative update, and the resulting *KABC–Second Edition Normative Update* (KABC–II NU; Kaufman & Kaufman, 2018a) has supplanted the previous version of the instrument. Although the test was re-normed on 500 participants, the test publisher declined to report updated structural validity results. Instead, it was argued in the manual supplement (Kaufman & Kaufman, 2018b) that because the content and organization of the test was not altered, users could consult the *KABC–II Technical Manual* (Kaufman, & Kaufman, 2004b) to infer the structure of the NU. After noting several anomalies in the manual supplement, McGill (2020) subjected the reported summary data to CFA and multidimensional scaling (see Figure 3.3). Results of the core battery analyses across school age clearly preferred a

hierarchical four-factor model, featuring a clean Perceptual Reasoning dimension. Models containing previously speculated cross-loadings were deemed inferior as were various bifactor model implementations, calling into question the viability of that model for the KABC–II NU (Bonifay et al., 2017).

<F03_003>

However, Reynolds and Keith (2017) argue that the Perceptual Reasoning consolidation is reductionistic and fails to consider that the identification problems are specific to Fluid Reasoning and not Visual Processing. Therefore an alternative model where Visual Processing is retained, and the Fluid Reasoning indicators load directly onto *g* should also be considered. Unfortunately, McGill (2020) did not examine this model. Consequently, we used the same summary data employed in that study and conducted a constrained CFA analysis using Mplus version 8.0 (Muthen & Muthen, 2017) with maximum likelihood estimation. The resulting fit statistics across school-age indicated that the model is statistically inferior to the final model retained by McGill (2020) and thus is not tenable for the data.

Luria Model Structural Validity

Structural investigations of the Luria model have been less pervasive in the literature. McGill and Spurgin (2017) conducted the first Luria EFA using the KABC–II normative data, and results did not support the publisher-preferred four-factor model. At ages 7–12, the Planning dimension was not mathematically viable, and Pattern Reasoning migrated to the Simultaneous Processing factor. At ages 13–18, Planning was again an impermissible factor and Pattern Reasoning did not load saliently on any factor. Figure 3.4 illustrates well the multidimensional nature of KABC–II measures and the dominance of *g*. Whereas general intelligence explained about 68% of subtest variance, no group-specific factors accounted for more than 16%. Although

indices of clinical relevance were not calculated, these variance estimates leave no doubt that the unique contributions of the first-order dimensions are likely to be minimal.

<F03_004>

McGill (2017) followed with an CFA investigation of the same data to further elucidate the Luria model structure. Results were more supportive of the four-factor publisher-preferred hierarchical four-factor model. However, the best fitting model required the specification of a Pattern Reasoning cross-loading on both Planning and Simultaneous Processing consistent with Reynolds et al. (2007). Based upon the procedures articulated in Reynolds and Keith (2013), an approximation of the SL procedure was also conducted, and the results did not support primary interpretation at the Stratum II index-level of the test.

Incremental Predictive Validity

Once the structure of an instrument has been firmly established,[7] it is then important to examine relationships with external measures—in particular, predictive validity (Wiggins, 1973). Given the dimensional nature of KABC–II scores, it is especially important to assess the incremental validity of lower-order CHC/Luria scores in predicting achievement beyond the FCI/MPI. In fact, the joint test standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) specifically require that when a test provides users with multiple scores, the distinctiveness of the scores be firmly established. As per Hunsley and Meyer (2003), McGill (2015) investigated the incremental validity of the CHC index scores in predicting achievement outcomes beyond that accounted for by the omnibus FCI score. Across the school-age range, $R^2/\Delta R^2$ effect size indices indicated that the FCI accounted for 30% to 65% of the variance across achievement outcomes. By contrast, the CHC scores jointly failed to account for more than 8% of the variance among

those measures. McGill and Spurgin (2016) replicated these results in their examination of the Luria interpretive model.

It is important to note that studies employing structural equation modeling as opposed to hierarchical multiple regression have furnished results largely supporting the findings reported above but have provided stronger evidence for the incremental validity of broad ability factors (e.g., Benson et al., 2016, Villeneuve et al., 2019). Nevertheless, in the latter study, Fluid Reasoning was again found to be redundant with the hierarchical general factor. Not surprisingly, a meta-analysis by Zaboski and colleagues (2018) examining the effects of CHC broad abilities on achievement (after controlling for *g*) found that the importance of the broad abilities was more circumspect and rarely did any individual dimension account for more than 10% of the variance in achievement.

Scatter and Diagnostic Utility

Perhaps the most controversial aspect of KABC–II assessment has been the procedures that have been advocated in various interpretive resources (e.g., Kaufman et al., 2005) for conducting scatter assessment amongst the hypothesized group-specific, factor-based indices. As previously noted, these guidelines do not appear to have been derived from any legitimate form of empirical evidence. Given this evidentiary lacuna, McGill (2016) conducted the first empirical examination of the potential iatrogenic impact of scatter on the structural or predictive validity of KABC–II scores. Put simply, no evidence was found to indicate that scatter had any meaningful effect on these outcomes. Subsequent investigations of the potential diagnostic value of scatter assessment on the KABC–II, which is a key principle undergirding emerging methods of specific learning disability (SLD) identification (i.e., PSW), have found that increasing levels of scatter predict SLD (McGill, 2018) and that intra-individual cognitive weaknesses discriminate between

individuals with and without achievement weaknesses at no better than chance levels (McGill et al., 2018). In sum, practitioners engaged in these assessment practices would likely be better served flipping coins as an alternative (Meehl & Rosen, 1966).

Implications of Dimensional Complexity for Clinical Interpretation

To be fair, the test authors have fully acknowledged the dimensional complexity of KABC–II measures insofar as that clinicians need to consider that many of the subtests do not provide a pure measurement of the hypothesized CHC/Luria abilities that they are purported to assess (Drozdick et al., 2018). In this way, cognitive complexity is integrated purposefully into core aspects of KABC–II design to enhance the purported clinical utility of the instrument. From a technical standpoint, such default statements are less problematic as researchers are able to disentangle effects using many of the previously described procedures for variance partitioning to determine the actual portions of variance explained by various psychological dimensions and the degree that measurement error may confound confidant clinical interpretation of that score. However, from a conceptual vantage point, practitioners on the ground are not able to account for these effects at the level of the individual and risk committing a misattribution error when attempting to extract meaning from the scores that they obtain from an examinee (Canivez, 2013). As previously discussed, several independent KABC–II CFA studies have found that Pattern Reasoning (which is theoretically assigned to Fluid Reasoning) loads on both Visual Processing and Fluid Reasoning, a finding that is not surprising given the issues occurred in trying to separate these factors from each other during the development of the instrument, suggesting there is tremendous overlap between those dimensions as conceptualized on the KABC–II.

It is certainly possible that the Pattern Reasoning cross-loading could be a methodological artifact and not the true measurement model underlying the data. In a Monte Carlo simulation study of the plausible models that have been reported in the literature for popular commercial ability measures, Dombrowski et al. (2021) illustrated well the danger in relying on any one CFA study as prima facie evidence for determining the actual structure of a test. Some speculative models evade replication when subsequently examined again even when using the same methodological parameters. In reconciling these discrepancies, it is important to consider the degrees of freedom afforded to researchers in CFA programs (Dombrowski & McGill, 2024; Waller & Meehl, 2002). Assuming that the Pattern Reasoning cross-loading is in fact legitimate, what are the implications for clinical interpretation? Like all IQ tests, the KABC–II employs a scoring structure that assumes orthogonality among its latent dimensions. Core subtests linearly combine to form one and only one first-order index score. When departures from desired simple structure are observed (e.g., theoretically inconsistent subtest migration, cross-loading, etc.) in external research studies, it calls into question the integrity of the scoring structure of the test as these anomalies are rarely, if ever, accounted for by test publishers.

In sum, the mathematical hurdles introduced during the developmental phase of an instrument cannot be overcome by skilled detective work no matter how many times the loaded term *clinical judgement* is invoked as an intellectual safety blanket to discount negative research findings (Lilienfeld & Strother, 2020; McGill et al., 2018). Despite the intuitive appeal of such sentiments, the misguided idea that a practitioner can turn specious assessment information into clinical gold has been regarded in the EBA literature as a form of alchemist's fallacy (Lilienfeld et al., 2006). The fact that such contentions continue to undergird much assessment training in

our business calls into question the idea that our field will ever engage in meaningful self-correction as a science.

Conclusion

In many respects the KABC–II/KABC–II NU is a fine instrument. Its materials and design at the time of its publication were innovative and its clinical versatility nonpareil. Nevertheless, as a result of the potential psychometric shortcomings noted in this chapter, a conflicting picture emerges as to how the instrument should be interpreted from an EBA perspective. On the one hand, there appears to be consensus that the KABC–II measures some of the hypothesized CHC dimensions that it purports to measure except for Fluid Reasoning and potentially Visual Processing. Whereas empirical evidence suggests that users can be reasonably confidant when interpreting the instrument at the global scale level (i.e., FCI, MPI, NVI), questions have been raised about the psychometric integrity of the first-order factors regardless of how they are theoretically conceptualized (i.e., CHC or Luria perspectives). This is not to say that practitioners should eschew interpretation of first-order factors as a matter of course. On the contrary, it is suggested that users engage in the interpretation of these indices with caution, considering unique contributions in measurement these indices afford beyond what is already known about the examinee from other available sources (Kranzler & Floyd, 2020). As compared to school age, limited examinations of preschool age structure support the more parsimonious interpretive model employed at that age range; however, more investigation is needed. It is unclear what, if any, clinical utility is afforded by the first-order indices with respect to treatment utility considerations.

Other matters discussed here appear to be more settled. For instance, the bulk of available research evidence does not provide support for the scatter assessment procedures that have

previously been advocated for the instrument (e.g., Kaufman et al., 2005). More importantly, despite the intuitive appeal of the dual theoretical model, it remains unclear how a score can mysteriously morph from a CHC construct to a Luria dimension simply because the examiner elects to use one interpretive scheme rather than another (Braden & Ouzts, 2005). Simply put, the omission of measures of Crystallized Ability do not suddenly convert the Short-Term Memory index into a measure of Sequential Processing. Accordingly, users of the KABC–II are cautioned against committing the naming fallacy (Borsboom, 2005) when electing to use the Luria interpretive model.

References

Beaujean, A. A. (2015). John Carroll's views on intelligence: Bi-factor vs. higher-order models. *Journal of Intelligence*, *3*(4), 121–136. https://doi.org/10.3390/jintelligence3040121

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

Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test use and assessment practices of school psychologists in the United States: Findings from the 2017 national survey. *Journal of School Psychology*, *72*, 29–48. https://doi.org/10.1016/j.jsp.2018.12.004

Benson, N. F., Kranzler, J. H., & Floyd, R. G. (2016). Examining the integrity of measurement of cognitive abilities in the prediction of achievement: Comparisons and contrasts across variables from higher-order and bifactor models. *Journal of School Psychology*, *58*, 1–19. https://doi.org/10.1016/j.jsp.2016.06.001

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science*, *5*(1), 184–186. https://doi.org/10.1177/2167702616657069

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511490026

Bracken, B. A. (1985). A critical review of the Kaufman Assessment Battery for Children (K–ABC). *School Psychology Review*, *14*(1), 21–36. https://doi.org/10.1080/02796015.1985.12085141

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). Buros Institute of Mental Measurements.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Byrne, B. M. (2005). Factor analytic models: Viewing the structure of an assessment instrument from three perspectives. *Journal of Personality Assessment*, *85*(1), 17–32. https://doi.org/10.1207/s15327752jpa8501_02

Caemmerer, J. M., Keith, T. Z., & Reynolds, M. R. (2020). Beyond individual intelligence test: Application of Cattell-Horn-Carroll theory. *Intelligence*, *79*, 101433. https://doi.org/10.1016/j.intell.2020.101433

Canivez, G. L. (2013). 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). Oxford University Press.

Canivez, G. L. (2016). Bifactor modeling in construct validation of multifactored tests: Implications for understanding multidimensional constructs and test interpretation. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advancements* (pp. 247–271). Hogrefe.

Carretta, T. R., & Ree, J. J. (2001). Pitfalls of ability research. *International Journal of Selection and Assessment*, *9*(1), 325–335. https://doi.org/10.1111/1468-2389.00184

Carroll, J. B. (1995). On methodology in the study of cognitive abilities. *Multivariate Behavioral Research*, *30*(3), 429–452. https://doi.org/10.1207/s15327906mbr3003_6

Dombrowski S. C. (2020) A newly proposed framework and a clarion call to improve practice. In S. C. Dombrowski (Ed.), *Psychoeducational assessment and report writing* (2nd ed., pp. 9–59). Springer. https://doi.org/10.1007/978-3-030-44641-3_2

Dombrowski, S. C., & McGill, R. J. (2024). Clinical assessment in school psychology: Impervious to scientific reform? *Canadian Journal of School Psychology*, *39*(4), 297–306. https://doi.org/10.1177/08295735231224052

Dombrowski, S. C., McGill, R. J., Canivez, G. L., Watkins, M. W., & Beaujean, A. A. (2021). Factor analysis and variance partitioning in intelligence test research: Clarifying misconceptions. *Journal of Psychoeducational Assessment*, *39*(1), 28–38. https://doi.org/10.1177/0734282920961952

Dombrowski, S. C., McGill, R. J., & Morgan, G. B. (2021). Monte Carlo modeling of contemporary intelligence test (IQ) factor structure: Implications for IQ assessment, interpretation, and theory. *Assessment*, *28*(3), 977–993. https://doi.org/10.1177/1073191119869828

Drozdick, L. W., Singer, J. K., Lichtenberger, E. O., Kaufman, J. C., Kaufman, A. S., & Kaufman, N. L. (2018). The Kaufman Assessment Battery for Children-Second Edition and KABC–II Normative Update. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 333–359). Guilford Press.

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

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*(5), 253–270. https://doi.org/10.1037/h0023816

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology*, *3*, 29–51. https://www.doi.org/10.1146/annurev.clinpsy.3.022806.091419

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment*, *15*(4), 446–455. https://doi.org/10.1037/1040-3590.15.4.446

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children*. Circle American Guidance Service.

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

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

Kaufman, A. S., & Kaufman, N. L. (2018a). *Kaufman Assessment Battery for Children-Second Edition Normative Update*. NCS Pearson.

Kaufman, A. S., & Kaufman, N. L. (2018b). *Kaufman Assessment Battery for Children-Second Edition Normative Update: Manual supplement*. NCS Pearson.

Kaufman, A. S., Kaufman, N. L., Drozdick, L. W., & Morrison, J. (2018). *Kaufman Assessment Battery for Children-Second Edition Normative Update manual supplement*. NCS Pearson.

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

Kaufman, J. C. (Ed.). (2009). *Intelligent testing: Integrating psychological theory and clinical practice*. Cambridge University Press.

Keith, T. Z. (1985). Questioning the K–ABC: What does it measure? *School Psychology Review*, *14*(1), 9–20, https://www.doi.org/10.1080/02796015.1985.12085140

Keith, T. Z., & Kranzler, J. H. (1999). The absence of structural fidelity precludes construct validity: Rejoinder to Naglieri on what the cognitive assessment system does and does not measure. *School Psychology Review*, *28*(2), 303–321. https://doi.org/10.1080/02796015.1999.12085967

Kranzler, J. H., & Floyd, R. G. (2020). *Assessing intelligence in children and adolescents: A practical guide for evidence-based assessment* (2nd ed.). Rowman and Littlefield.

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology*, *61*(4), 281–288. https://doi.org/10.1037/cap0000236

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2007). Why questionable psychological tests remain popular. *Scientific Review of Alternative Medicine*, *10*, 6–15.

Luria, A. R. (1966). *Human brain and psychological processes*. Harper Row.

McGill, R. J. (2015). Interpretation of KABC–II scores: An evaluation of the incremental validity of CHC factor scores in predicting achievement. *Psychological Assessment*, *27*(4), 1417–1426. https://doi.org/10.1037/pas0000127

McGill, R. J. (2016). Invalidating the full scale IQ score in the presence of significant factor score variability: Clinical acumen or clinical illusion? *Archives of Assessment Psychology*, *6*(1), 49–79.

McGill, R. J. (2017). Exploring the latent structure of the Luria model for the KABC–II at school age: Further insights from confirmatory factor analysis. *Psychology in the Schools*, *54*(9), 1004–1018. https://doi.org/10.1002/pits.22037

McGill, R. J. (2018). Confronting the base rate problem: More ups and downs for cognitive scatter analysis. *Contemporary School Psychology*, *22*(3), 384–393. https://doi.org/10.1007/s40688–017-0168-4

McGill, R. J. (2020). An instrument in search of a theory: Structural validity of the Kaufman Assessment Battery for Children-Second Edition Normative Update at school-age. *Psychology in the Schools*, *57*(2), 247–264. https://doi.org/10.1002/pits.22304

McGill, R. J., & Dombrowski, S. C. (2018). Factor structure of the CHC model for the KABC–II: Exploratory factor analyses with the 16 core and supplementary subtests. *Contemporary School Psychology*, *22*(3), 279–293. https://doi.org/10.1007/s40688-017-0152-z

McGill, R. J., Dombrowski, S. C., & Canivez, G. L. (2018). Cognitive profile analysis in school psychology: History, issues, and continued concerns. *Journal of School Psychology*, *71*, 108–121. https://doi.org/10.1016/j.jsp.2018.10.007

McGill, R. J., & Spurgin, A. R. (2016). Assessing the incremental value of KABC–II Luria model scores in predicting achievement: What do they tell us beyond the MPI? *Psychology in the Schools*, *53*(7), 677–689. https://doi.org/10.1002/pits.21940

McGill, R. J., & Spurgin, A. R. (2017). Exploratory higher order analysis of the Luria interpretive model on the Kaufman Assessment Battery for Children-Second Edition (KABC–II) school-age battery. *Assessment*, *24*(4), 540–552. https://doi.org/10.1177/1073191115614081

McGrew, K. S., LaForte, E. M., & Shrank, F. A. (2014). *Woodcock–Johnson IV: Technical Manual.* Riverside Publishing.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*(3), 194–216. https://doi.org/10.1037/h0048070

Muthen, L. K., & Muthen, B. O. (2017). *Mplus* [Version 8.0]. Muthen & Muthen.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/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*(2), 129–140. https://www.doi.org/10.1080/00223891.2012.725437

Reynolds, M. R., & Keith, T. Z. (2013). Measurement and statistical issues in child assessment research. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwean (Eds.), *The Oxford handbook of child psychological assessment* (pp. 48–83). Oxford University Press.

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

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

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

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3), 223–237. https://doi.org/10.1080/00223891.2015.1089249

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

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll Theory of Cognitive Abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73–163). Guilford Press.

Sotelo-Dynega, M., & Dixon, S. G. (2014). Cognitive assessment practices: A survey of school psychologists. *Psychology in the Schools*, *51*(10), 1031–1045. https://doi.org/10.1002/pits.21802

Villeneuve, E. F., Hajovsky, D. B., Mason, B. A., & Lewno, B. M. (2019). Cognitive ability and math computation developmental relations with math problem solving: An integrated, multigroup approach. *School Psychology Quarterly*, *34*(1), 96–108. https://doi.org/10.1037/spq0000267

Waller, N. G., & Meehl, P. E. (2002). Risky tests, verisimilitude, and path analysis. *Psychological Methods*, *7*(3), 323–337. https://doi.org/10.1037/1082-989X.7.3.323

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Addison-Wesley.

Youngstrom, E. A. (2013). Future directions in psychological assessment: Combining evidence-based medicine innovations with psychology's historical strengths to enhance utility. *Journal of Clinical Child & Adolescent Psychology*, *42*(1), 139–159. https://www.doi.org/10.1080/15374416.2012.736358

Zaboski, B. A., Kranzler, J. H., & Gage, N. A. (2018). Meta-analysis of the relationships between academic achievement and broad abilities of the Cattell-Horn-Carroll theory. *Journal of School Psychology*, *71*, 42–56. https://www.doi.org/10.1016/j.jsp.2018.10.001

Figure 1

Hypothesized KABC-II CHC Measurement Model

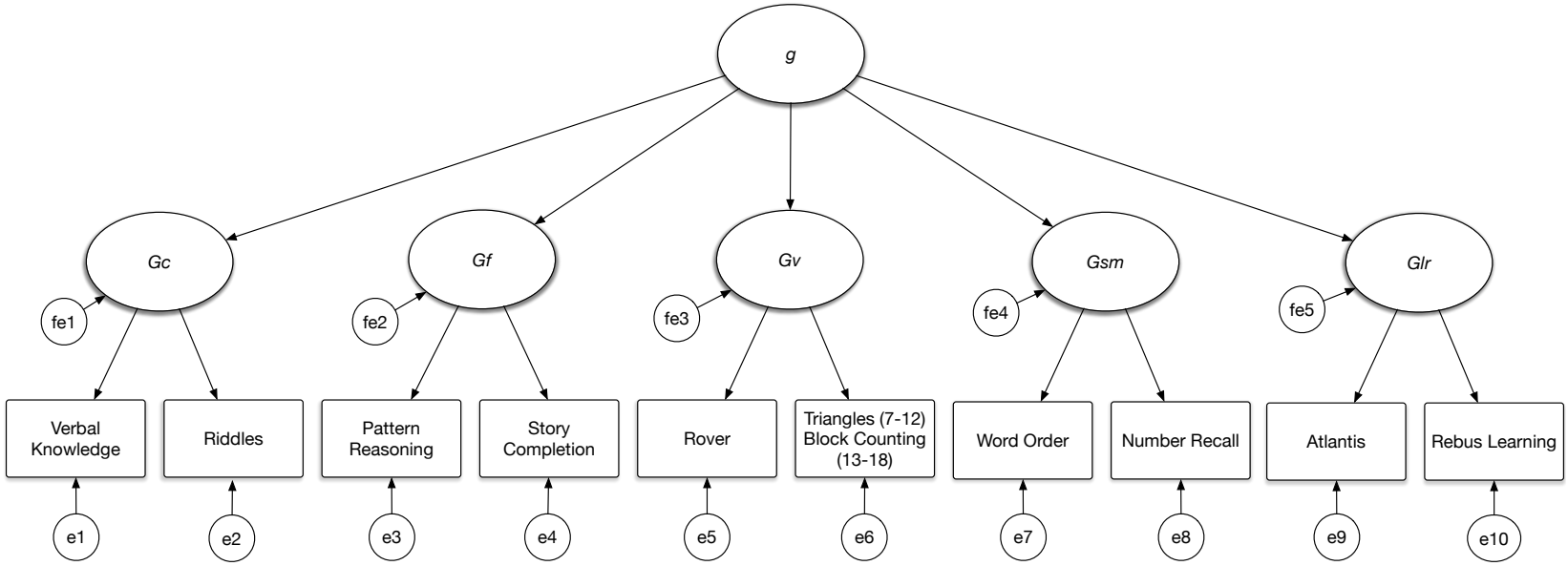


Figure 2

Hypothesized KABC-II Luria Measurement Model

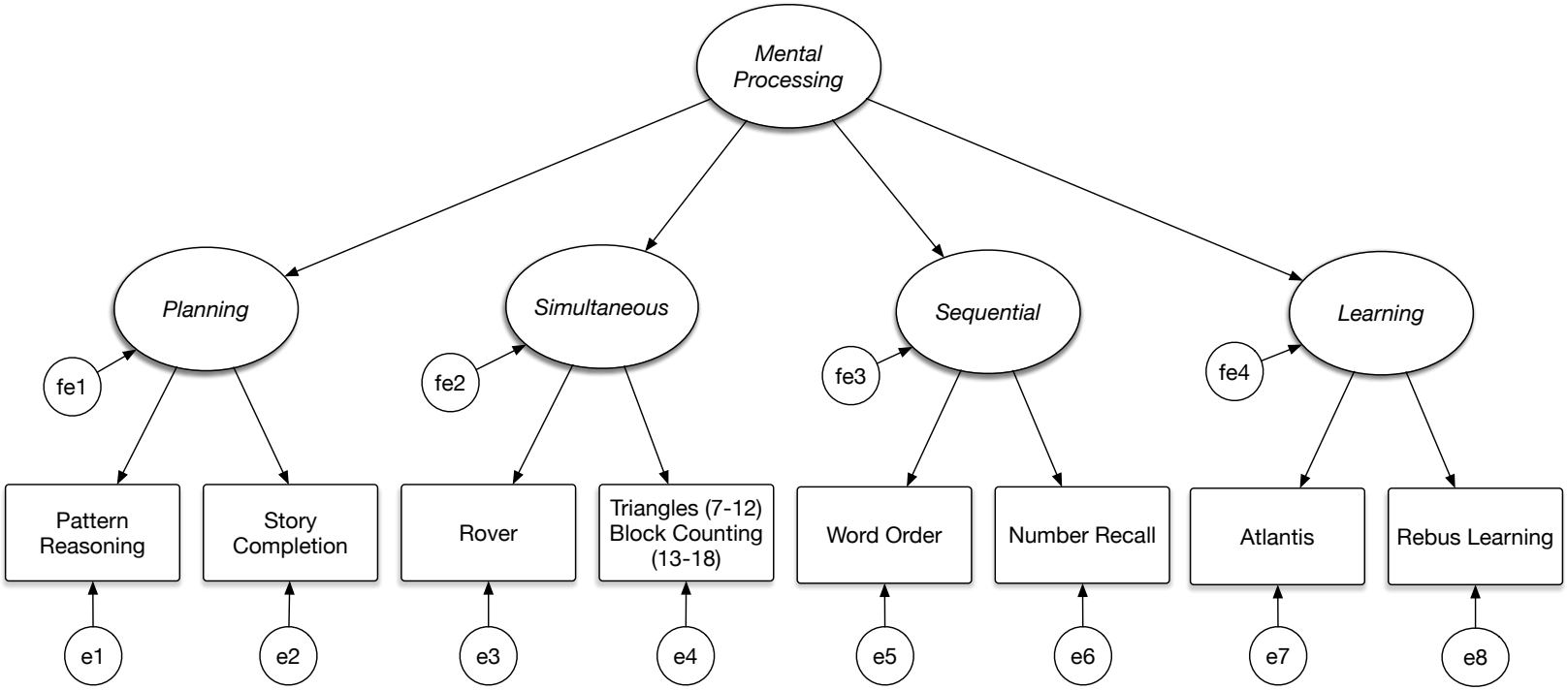


Figure 3

Radex Model of the KABC-II CHC Core Battery for ages 13-18

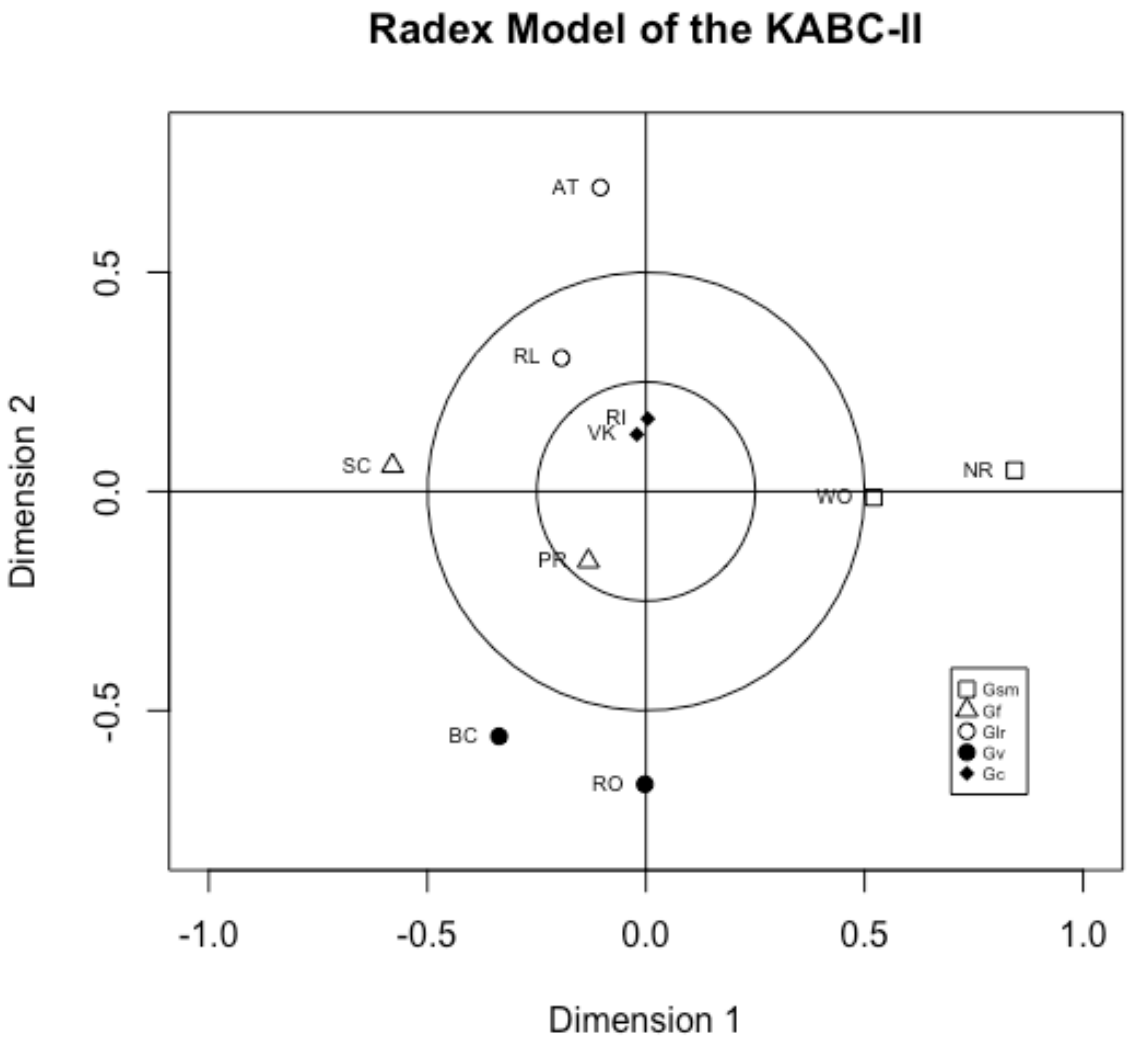
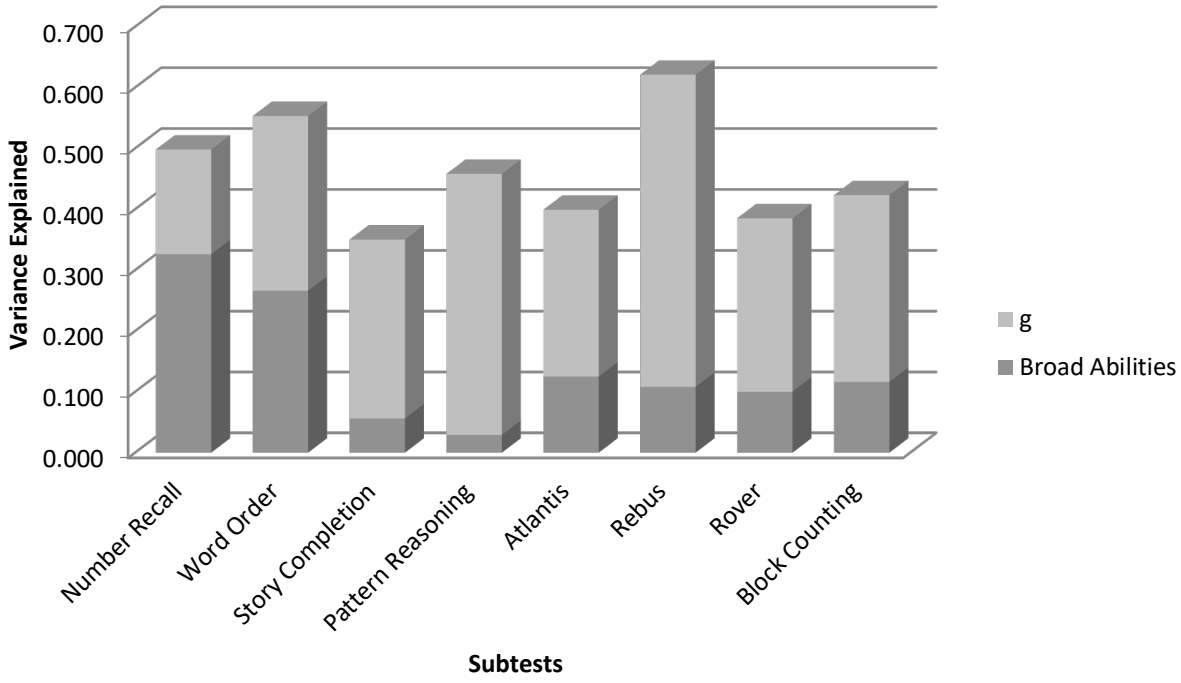


Figure 4

Sources of Variance in KABC-II Luria Model Subtests



[1] Elements of this chapter were presented at the 2019 meeting of the National Association of School Psychologists, Atlanta, Georgia.

[2] Specific subtests can only be administered at restricted age ranges and may not be administered across batteries.

[3] The Visual Processing Index is produced from a different combination of subtests at different ages. Whereas Rover remains constant, Triangles is used at ages 7–12, and Block Counting is used at ages 13–18.

[4] The KABC–II has been adapted and translated for use in several additional countries.

[5] For examinees with limited verbal ability in English, an ancillary Nonverbal Index (NVI) is available using different combinations of subtests at different points in the age range. This index is thought to serve as an alternative estimate of general intelligence (*g*).

[6] It is possible to use CFA programs in an exploratory manner, however, the number of closed-door analyses that appear to have been run raise concern as to whether the models that were ultimately retained for the instrument were obtained through capitalizing on chance (Byrne, 2005).

[7] We would argue that there remain significant questions as to what the KABC–II/KABC–II NU measures.