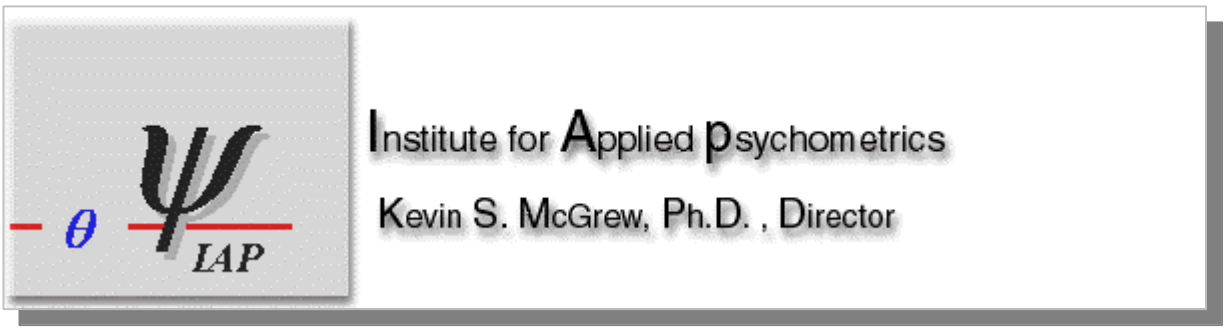


Research Report No. 1

The Measurement of Reading Achievement by Different Individually Administered Standardized Reading Tests: Apples and Apples, or Apples and Oranges?

Kevin S. McGrew, Ph.D.
Director

Institute for Applied Psychometrics llc
Clearwater, MN

This report was originally written for the *Committee on the Prevention of Reading Difficulties in Young Children* as a background paper for: Snow, C. E., Burns, S. M., & Griffin, P.E (Eds) (1998), *Preventing reading difficulties in young children*, Washington: National Academy Press.

ACKNOWLEDGEMENTS

The author would like to thank American Guidance Services, Riverside Publishing, and the Psychological Corporation for providing copies of the tests that were reviewed. Recognition is also extended to Mark Daniel, Dawn Flanagan, Jo Heille, Gary Hessler, Nancy Mather, and Richard Woodcock. These individuals facilitated in the preparation of this manuscript by providing relevant information (i.e., references, articles, data) and/or by commenting on earlier versions of the paper.

What is reading but silent conversation?
Walter Savage Landor, "Arisoteles and Callisthenes",
Imaginary Conversations (1823-53)

INTRODUCTION AND PURPOSE

Reading is largely an unobservable mental activity. As a result, reading ability must be inferred indirectly from observable behavior. To accomplish this function, a variety of product and process approaches to the measurement of reading have been developed (Johnston, 1983). Of these, the predominant technology that has been employed in reading placement, classification, monitoring, and evaluation activities is *standardized achievement testing*. This approach to measurement relies on the use of standardized stimuli and procedures to elicit observable responses that represent the "products" of an examinee's learning (Johnston, 1983). These observable performances are then compared against a standard (i.e., a norm group) in order to quantify an individual's level of performance.

Two general categories of standardized reading tests are typically used in the measurement of reading achievement (Joshi, 1995). Reading *screening*, *survey* or *general achievement* tests are used to measure what and how much an individual has learned (i.e., an individual's relative level of performance), while *diagnostic* reading tests are designed to identify in detail an individual's relative strengths and weaknesses in reading, with an eye toward the development of a specific intervention program (Joshi, 1995; Salvia & Ysseldyke, 1991). The general reading achievement tests are those that are most frequently used when making classification and placement decisions and when measuring outcomes in reading disability program evaluation and research studies.

Since standardized reading achievement tests provide much of the data in the reading research literature and thus, may influence critical research and policy decisions, it is imperative that the tests used to produce these data be technically sound. Also, in order to draw accurate and meaningful conclusions from the vast array of reading disability research, it is important that individuals who integrate this research understand the degree to which the data from different standardized reading tests are comparable (i.e., measure the same constructs). One cannot always assume that two tests that are similarly named (e.g., reading comprehension) are measuring identical reading constructs.

The purpose of this paper is to provide a better understanding of the most frequently used *individually administered standardized reading achievement tests* (hereafter referred to as reading achievement tests).[1] The primary goals of this paper are to:

- describe and compare the characteristics of the most frequently used reading achievement tests,
- evaluate the extent to which the most frequently used reading achievement tests possess adequate psychometric characteristics, and
- identify the common and uncommon cognitive constructs measured by the most frequently used reading achievement tests.

[1]Group standardized reading achievement tests are not treated in this paper. Group achievement tests differ from individual tests in a number of salient areas (e.g., the length of testing time for the examinee is longer in a group test; group tests rely extensively on a multiple choice response format; etc.)

Finally, at the request of the *NAS Committee on the Prevention of Reading Disabilities in Young Children*, this paper will also provide an evaluation and discussion of an approach to measuring reading comprehension that is often questioned. The validity of the *modified cloze* approach to measuring reading comprehension will be examined.

DESCRIPTION, COMPARISON, AND EVALUATION OF FREQUENTLY USED READING ACHIEVEMENT TESTS

More than a dozen individually administered general achievement batteries exist, but only approximately half of these are used extensively (Gregory, 1996). Surveys (Harrison, Kaufman, Hickman, & Kaufman, 1988; Mardell-Czudnowski, 1996; Stinnett, Havey & Oehler-Stinnett, 1994) of a variety of assessment personnel consistently report that the *Kaufman Assessment Battery for Children* (K-ABC; Kaufman & Kaufman, 1983), *the Kaufman Test of Educational Achievement* (K-TEA; Kaufman & Kaufman, 1985), *the Peabody Individual Achievement Test--Revised* (PIAT-R; Markwardt, 1989), the various editions of the *Wide Range Achievement Test* (WRAT-3; Wilkinson, 1993), and the achievement section of the *Woodcock-Johnson Psychoeducational Battery--Revised* (WJ-R; Woodcock & Johnson, 1989) are the most frequently used individually administered standardized achievement batteries.

Three relatively new achievement batteries are likely to experience similar popularity. By nature of it's statistical link to the most frequently used children's intelligence test battery (viz., the WISC-III), the *Wechsler Individual Achievement Test* (WIAT; The Psychological Corporation, 1992) is likely to be a frequently used test battery. By nature of their shared content and common norm sample with the popular WJ-R, the *Woodcock-McGrew-Werder Mini-Battery of Achievement* (MBA; Woodcock, McGrew, & Werder, 1994) and the *Woodcock Diagnostic Reading Battery* (WDRB; Woodcock, 1997) will also probably become frequently used.

The achievement batteries listed above are likely to be the most frequently used individually administered standardized batteries in psychoeducational contexts. These batteries are the focus of the remainder of this paper. A general description of the reading tests in each of these batteries follows below. Included in the test descriptions are general evaluative judgments regarding the overall technical adequacy of each test. These evaluations are based on published independent test reviews. A summary and comparison of the reading tests follows these descriptions.

Kaufman Assessment Battery for Children (K-ABC)

The K-ABC is a battery of cognitive and achievement tests that was published in 1983. The battery was standardized on a nationally representative sample of 2,000 subjects from ages two through twelve (Kaufman & Kaufman, 1983). The achievement battery includes tests that tap the domains of reading, mathematics, and general knowledge. The achievement battery includes two reading tests:

- Reading Decoding. This test measures an individual's ability to recognize and pronounce letters and words in a nonmeaningful context. The examinee is not required to know the meaning of any of the words.
- Reading Understanding. This test measures reading comprehension. The examinee is required to read sentences and then "act out" or perform the actions conveyed by the sentences.

Reviewers of the K-ABC have concluded that the "measures have sound psychometric qualities" (Coffman, 1985, p. 772) and that it meets "high standards of technical quality" (Anastasi, 1985, p. 771). *The K-ABC reading tests appear to meet established standards for measuring and evaluating the reading achievement levels of individuals.*

Kaufman Test of Educational Achievement (K-TEA)

The K-TEA is a battery of achievement tests that was published in 1985 (Kaufman & Kaufman, 1985). The battery has "brief" and "comprehensive" forms, with the Comprehensive form being the focus of this paper. The battery was standardized on a nationally representative sample of 1, 409 (spring) and 1,067 (fall) individuals in grades 1 through 12.[2] The achievement section of the battery includes tests in the areas of reading, spelling, and math.

The K-TEA includes two individual reading tests. These two tests can be combined into the Reading Composite score. The two K-TEA reading tests are:

- Reading Decoding: This subtest requires examinee's to identify letters and pronounce words of gradually increasing difficulty, both phonetic and nonphonetic. The words are presented in a nonmeaningful context and the examinee is not required to know the meaning of the words.
- Reading Comprehension: The test measures reading comprehension by requiring examinee's to read passages. The examinee must then read and orally answer one or two literal and/or inferential comprehension questions about the passages (which stay in view of the subject while they answer the questions). The easiest and most difficult items in the test use a different format that require the examinee to respond either gesturally (easy items) or orally (difficult items) to commands that are given in printed sentences.

The K-TEA appears to provide psychometricaly sound measurement of reading skills (Doll, 1989; Lewandowski, 1988). The K-TEA has been described as a "well-standardized, reliable measure" (Doll, 1989, p. 412). *The K-TEA reading tests appear to meet established standards for measuring and evaluating the reading performance levels of individuals.*

Peabody Individual Achievement Test--Revised (PIAT-R)

The PIAT-R is a 1989 revision (Markwardt, 1989) of a battery of achievement tests that was first published in 1970. The battery was standardized on a nationally representative sample of 1, 563

[2] New norms will be published in the fall of 1997 (Daniel, 1997).

kindergarten through 12th grade students.[3] The achievement section of the battery includes tests in the areas of reading, writing, math, and general information.

The PIAT-R includes two individual reading tests that can be combined into a Total Reading score. The two PIAT-R reading tests are:

- Reading Recognition: The test measures an examinee's skills in translating printed letters or words into speech sounds. The easiest items assess letter identification and knowledge of beginning sounds through multiple choice items. The next series of items require the subject to read a word and then select the correct picture that represents the word (a multiple choice format). Subsequent items require the examinee to read lists of individual words in a nonmeaningful context. The examinee is not required to know the meaning of the words.
- Reading Comprehension: This test is designed to measure an examinee's ability to derive meaning from printed words. The test uses a multiple choice format that requires the examinee to silently read a single sentence, and then after the sentence is removed, select from four pictures the one picture that best represents the meaning of the sentence.

Reviews (Allinder & Fuchs, 1992; Bennes, 1992; Joshi, 1995; Luther, 1992) of the PIAT-R are in general agreement that this battery is a technically adequate instrument that "appears to be a useful instrument to both practitioners in the schools and researchers" (Rogers, 1992, p.652). *The PIAT-R reading tests appear to meet established standards for measuring and evaluating the reading achievement levels of individuals.*

Wechsler Individual Achievement Test (WIAT)

The WIAT is a battery of achievement tests in reading, writing, mathematics, and language (i.e., oral expression and listening comprehension) that was published in 1992. The battery was standardized on a nationally representative sample of 4,252 kindergarten through 12th grade subjects (The Psychological Corporation, 1992). The battery consists of two reading tests called Basic Reading and Reading Comprehension that can be combined into a Reading Composite score. The two WIAT reading tests are:

- Basic Reading: This test uses a series of pictures and printed words to assess decoding and word recognition or identification skills. Early items require the examinee to point while later items require an oral response. The examinee is presented the words in a non-meaningful context and is not required to know the meaning of the words.
- Reading Comprehension: This test is designed to measure reading comprehension by requiring the examinee to respond orally to literal and inferential comprehension questions after reading passages (which stay in view of the subject while they answer the questions).

[3] New norms will be published in the fall of 1997 (Daniel, 1997).

Reviewers have consistently concluded that the WIAT is a psychometrically sound instrument (Joshi, 1995; Nicholson, 1992; Thompson, 1993). *The WIAT reading tests appear to meet established standards for measuring and evaluating the reading achievement levels of individuals.*

Wide Range Achievement Test: Third Edition (WRAT-3)

The WRAT-3 is the seventh edition of three brief screening tests in reading, spelling, and arithmetic that was first published in 1936. The current version was standardized on a national sample of 4,433 subjects from 5 through 74 years of age (Wilkinson, 1993). The test includes one measure of reading:

- Reading: The examinee is required to recognize and pronounce words in isolation (i.e., not in a meaningful context). The examinee is not required to know the meaning of any of the words.

Despite being a frequently used measure of reading, independent reviews of most editions of the WRAT have not been favorable. For example, after reviewing the sixth edition (i.e., WRAT-R), Harrison (1989) concluded: "does the WRAT-R meet psychometric standards for an achievement test? This reviewer must answer 'No' given available information" (p. 905). Harrison had concerns about the description of the standardization sample, the correct application of the Rasch model to item scaling, and questionable reliability and content and concurrent validity. Unfortunately, Clark's (1992) conclusion that "perhaps the worst thing that can be said about the WRAT-R is that the more it changes, the more it stays the same" appears to have been prophetic about the seventh edition (i.e., WRAT-3). Mabry (1995) and Ward (1995) both consider the WRAT-3 to have questionable psychometric characteristics. Mabry (1995) concluded that "[s]uspicions of inadequacy...are confirmed by test content, item formats, obsolescence, underlying philosophy, potential for bias, and insufficient evidence of validity and reliability" (p.1110). *Independent reviews indicate that the WRAT-3 Reading test does not meet acceptable psychometric standards for use as a measure of reading achievement.*

Woodcock-Johnson Psychoeducational Battery--Revised (WJ-R) and Woodcock Diagnostic Reading Battery (WDRB)

The WJ-R is a comprehensive battery of cognitive and achievement tests that was published in 1989 (Woodcock & Johnson, 1989). The battery was standardized on a nationally representative sample of 6,359 subjects that ranged in age from two years to late adulthood (McGrew, Werder, & Woodcock, 1991). The achievement section of the battery includes tests in the areas of reading, writing, math, and general knowledge. The WJ-R includes four individual reading tests that can be combined in three different two-test clusters (viz., Basic Reading Skills, Reading Comprehension, Broad Reading), the recommended level of interpretation by the test author (in order to increase the reliability and validity of interpretations). [The WDRB includes the exact same four reading tests (as well as select WJ-R cognitive tests) that are included in the WJ-R. Thus, comments made about the WJ-R reading tests are also applicable to the WDRB. This will be reflected throughout the remainder of this paper by referring to these tests as WJ-R/WDRB]. The four tests are:

- Letter-Word Identification: This test measures an examinee's reading identification skills. The first five items require the examinee to match a rebus (pictographic representation of a word) with an actual picture of the object. The remaining items require the identification of isolated letters and words. The examinee is not required to know the meaning of any of the words.
- Word Attack: The Word Attack test measures an examinee's ability to apply phonic and structural analysis skills to the pronunciation of unfamiliar printed words. The examinee must read aloud letter combinations that, while linguistically logical in English, form nonsense or low-frequency words in the English language.
- Reading Vocabulary: Reading Vocabulary measures the examinee's skill in reading words and supplying appropriate meanings. The test requires examinees to read words and then state a word that is either similar (synonyms) or opposite (antonyms) in meaning to the word presented.
- Passage Comprehension: This test is designed to measure reading comprehension. For the first four items, reading comprehension is measured by requiring the examine to point to the picture represented by a phrase (a multiple choice format). The remaining items require examinees to read short passages and to identify a missing key word in the passage. This task requires the examinee to state a word that would be appropriate in the context of the passage. This is an example of a *modified cloze* procedure.

Reviews of the technical adequacy of the WJ-R/WDRB reading tests, as well as reviews of the original edition of the WJ reading tests and the "sister" Woodcock Reading Mastery Test (WRMT-R; Woodcock, 1987), have all been positive (Cooter, 1989; Jaeger, 1989; Joshi, 1995; Kaufman, 1985; Salvia & Ysseldyke, 1991). The WJ/WJ-R/WDRB has been described as a "technically excellent instrument with exceptional reliability and concurrent validity" (Kaufman, 1985). *The WJ-R/WDRB reading tests appear to meet established standards for measuring and evaluating the reading achievement levels of individuals.*

Woodcock-McGrew-Werder Mini-Battery of Achievement (MBA)

The MBA is a battery of reading, writing, math, and factual knowledge tests published in 1994 (Woodcock et al,, 1994). The MBA provides a single total Reading score. The MBA reading score is described below.

- Reading. The total Reading cluster is a combination of the Identification, Reading Vocabulary, and Comprehension subtests. The score reflects the combination of an examinee's ability to (a) identify isolated letters and words out of context, (b) read words and provide a word opposite in meaning to the word presented, and (c) read a short passage and identify a missing word (i.e., a modified cloze procedure).

Given the common items, shared standardization sample, and test development procedures with the WJ-R/WDRB, the reviews listed above for the WJ-R/WDRB hold true for the MBA. An separate

review of the MBA (Cruse, Dumont & Willis, 1996) also concluded that the MBA is a technically sound instrument. *The MBA reading test appears to meet established standards for measuring and evaluating the reading performance achievement of individuals.*

Summary And Comparison Of Reading Achievement Tests

A review of the selected individually administered standardized achievement batteries that include tests of reading reveal both similarities and differences.

Breadth of Reading Skills Measured

The breadth of reading skills assessed by the different reading tests ranges from a singular narrow focus on word identification skills (WRAT-3) to a broad focus on word identification, word attack, reading vocabulary and comprehension (WJ-R/WDRB). Although it only provides a single reading score, the MBA is next to the WJ-R/WDRB in the breadth of skills sampled (i.e., it measures all the same domains as the WJ-R/WDRB except word attack). The K-TEA, K-ABC, PIAT-R, and WIAT tests fall in the middle as all measure word identification and reading comprehension. With the exception of the WRAT-3, all batteries include measures of both word identification and reading comprehension.

Assessment of Word Identification

With one exception, all seven reading tests reviewed assess word identification skills in a similar manner. Although there are slight differences in the format and procedures used for the easiest items (i.e., those dealing with letter names and sounds), all the reading tests include a test that assesses word identification skills through the presentation of lists of words in a non-meaningful context. The WJ-R/WDRB is unique in that it includes an additional test (viz., Word Attack) that assesses application of phonics and structural analysis skills to words.

Assessment of Comprehension

"Attempting to define reading comprehension is always a bold move--dissenters are likely to be many" (Johnston, 1983, p.1). Not unexpectedly, a variety of different approaches have been developed to assess reading comprehension. This variety is present in the reading comprehension tests included in this review. Standard test item terminology (Osterlind, 1989) can be used to describe and classify the different reading comprehension tests.

The area of greatest similarity across tests is in *item length*. With two exceptions, all the reading comprehension tests require examinee's to read sentences and/or short passages. The PIAT-R differs from the other tests in that it only uses single sentence item types and the WJ-R/WDRB Reading Vocabulary tests require examinees to read isolated words (i.e., not connected discourse).

The *mode of response* required by the examinee is the next area of greatest communality. The comprehension tests of the K-TEA , WIAT, WJ-R/WDRB, and MBA all require an oral response by

the examinee. However, the K-TEA differs from the WIAT, WJ-R/WDRB, and MBA by requiring a gestural response from examinees on the test's easy and difficult items. The K-ABC and PIAT-R are the primary "outliers" on the mode of response dimension. The K-ABC Reading Understanding test requires gestural responses. In the case of the PIAT-R, although an examinee can respond with a one-word oral response, all that is required is for the examinee to point.

Item format is the area of greatest difference across the reading comprehension tests. The PIAT-R Reading Comprehension test is unique in the use of multiple choice items, an example of a *selected-response* format. With the exception of the K-TEA Reading Comprehension test, all other comprehension tests use some form of *constructed-response* format However, the specific constructed-response format used does vary across the comprehension tests.

The WJ-R/WDRB Passage Comprehension and part of the MBA Reading test both assess reading comprehension via the *modified cloze* procedure. In addition, the WJ-R/WDRB Reading Vocabulary test and part of the MBA Reading test employes an *oral short-answer* format. The WIAT Reading Comprehension test stands alone in the use of an *oral open-ended* format, while the K-ABC Reading Understanding test is unique in the use of a *gestural open-ended* format. Finally, the K-TEA Reading Comprehension test is unique in that it includes a mixture of item formats. Different K-TEA items require examinee's to provide either *gestural or oral open-ended* responses (i.e., constructed-responses). The K-TEA Reading Comprehension test also includes items that use the *multiple choice* (i.e., selected-response) format.

Given the variability in the item format and response modes of the different reading comprehension tests, especially when compared to the relatively high degree of homogeneity of format and response mode on the various word identification tests, one would expect that the intercorrelations among reading comprehension tests (which is one indicator of construct validity) would be less than that observed for the word identification tests. Empirical data regarding this issue are presented later.

Technical Adequacy

The consensus of independent reviewers of the K-ABC, K-TEA, MBA, PIAT-R, WIAT, and WJ-R/WDRB is that all of these test batteries include individual reading tests that possess adequate psychometric characteristics for use in the measurement and evaluation of an individual's level of reading performance. This is not to say that reviewers have not raised other concerns (e.g., administration procedures, complexity of scoring, use of a particular response format, fall and spring norms vs continuous year norms, etc.). Also, this does not mean that these tests necessarily possess adequate validity for use as diagnostic reading tests. *The reading tests included in the K-ABC, K-TEA, MBA, PIAT-R, WIAT, and WJ-R/WDRB all appear to meet established standards for measuring and evaluating the reading achievement levels of individuals.*

Unfortunately, the WRAT-3 retains much of its lineage with prior editions, editions for which reviewers have consistently found technical faults. Based on its questionable technical characteristics and limited range of reading skills, it is recommended that the *WRAT-3 reading test not be used in the evaluation of individuals or groups when important decisions are being made (e.g., classification, placement, monitoring progress, program evaluation). The results of reading studies that use the*

WRAT-3 to classify and describe, monitor progress, and/or evaluate the effectiveness of intervention programs should be viewed with caution.

THE NEED FOR A ROADMAP: A CONCEPTUAL TAXONOMY FOR EVALUATING THE CONSTRUCTS MEASURED BY PSYCHOEDUCATIONAL TESTS

During the past decade increased attention has focused on multidimensional theories and models of human cognitive abilities. Some of the most prominent are Carroll's Three-Stratum Theory of Cognitive Abilities, Gardner's Theory of Multiple Intelligenceís, the Horn-Cattell Fluid-Crystallized (*Gf-Gc*) Theory of Intelligence, the Luria-Das Model of Information Processing, and Sternberg's Triarchic Theory of Intelligence (see Flanagan, Genshaft, & Harrison, 1997, for a review). Of these theories, the Horn-Cattell *Gf-Gc* Theory (Horn, 1991, 1994; Horn & Noll, 1997) and the Three-Stratum Theory of Cognitive Abilities (Carroll, 1993, 1997) have been found to be the most comprehensive and extensively researched *psychometric* frameworks from which to evaluate the abilities measured by psychoeducational assessment instruments (Flanagan & McGrew, 1997; McGrew & Flanagan, 1996; McGrew, Flanagan, Keith, & Vanderwood, in press; Messick, 1992).

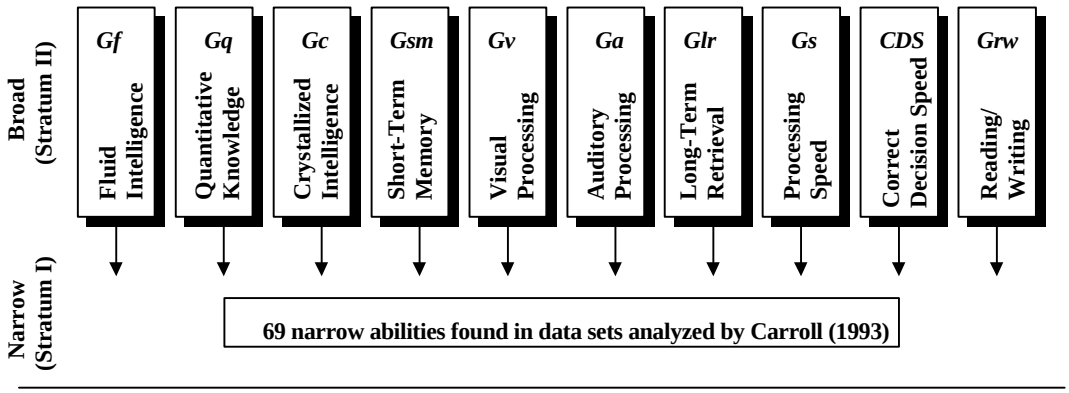
The *Gf-Gc* Cognitive Ability Model

Reviews of the factor analytic research on human abilities (Carroll, 1993; Gustafsson, 1984, 1988; Horn, 1988, 1991, 1994; Lohman, 1989), and Carroll's (1993) seminal review of the extant factor analytic research in particular, have converged on the *Gf-Gc* multiple cognitive abilities framework that underlies both the Carroll and Horn-Cattell models. Given the breadth of empirical support for the *Gf-Gc* structure of human cognitive abilities, models based on the *Gf-Gc* theory (i.e., the Horn-Cattell and Carroll models) currently provide the most useful psychometricaly based framework for identifying the constructs measured by psychoeducational assessment instruments (McGrew, 1997). Other frameworks or "lenses" (e.g., developmental, cognitive or information processing; sociocultural) from which to view reading assessment (Palincsar & Perry, 1995) are not discussed in this paper given that the questions that need to be addressed here are psychometric in nature.

It is not possible to adequately describe the Horn-Cattell and Carroll *Gf-Gc* based models in sufficient detail in this paper. The reader is referred to Horn (1991, 1994), Horn and Noll (1997), and Carroll (1993, 1997) for a comprehensive description of these theoretical frameworks. For this paper, the synthesized Horn-Cattell and Carroll *Gf-Gc* model outlined by McGrew (1997), the model that was recently used to map the individual tests in all the major individually administered intelligence test batteries to the *Gf-Gc* taxonomy (McGrew, 1997; McGrew & Flanagan, 1997), will be used to identify and compare the abilities measured by reading achievement tests. This model is presented in Figure 1.

A central feature of the model in Figure 1 is the understanding that human cognitive abilities can be organized hierarchically by levels of breadth. The model presented in Figure 1 indicates

that there are at least 10 *broad* (stratum II*) Gf-Gc* abilities which subsume a large number of *narrow* (stratum I) abilities. Evaluating and comparing individual reading achievement tests requires the identification of the *Gf-Gc* cognitive abilities that contribute to performance on each test.



Fluid Intelligence (*Gf*) - the ability to reason (inductive and deductive) and solve novel problems.
Quantitative Knowledge *(Gq)* - store of acquired quantitative information and knowledge (e.g., concepts, vocabulary).
Crystallized Intelligence (*Gc*) - breadth and depth of knowledge about a culture (e.g., general information, vocabulary).
Short-Term Memory (*Gsm*) - the ability to hold information in immediate awareness.
Visual Processing *(Gv)* - the ability to analyze, synthesize, and manipulate visual images in space.
Auditory Processing *(Ga)* - the ability to analyze, synthesize, and manipulate auditory stimuli.
Long-Term Associate Storage and Retrieval *(Glr)* - the ability to store and then later retrieve information through association.
Processing Speed *(Gs)* - the ability to rapidly perform simple cognitive tasks automatically for a sustained period of time.
Correct Decision Speed *(CDS)* - speed in providing quick answers to questions/problems.
Reading/Writing (*Grw)* - basic and complex comprehension and expression skills in reading and writing.

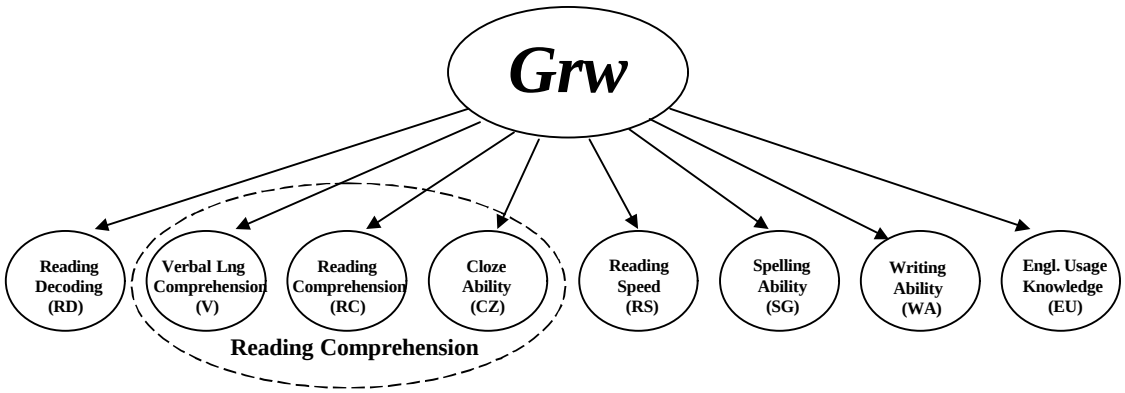
Figure 1. Horn-Cattell *Gf-Gc* Broad Ability Model and Definitions

Gf-Gc Abilities Relevant to Reading

The *Grw* (reading/writing) domain represents a broad ability associated with basic and complex comprehension and expression skills in reading and writing (Horn, 1988; McGrew et al., 1991; Woodcock, in press). *Grw* subsumes a number of narrow reading and writing factors that have been identified within the English language (Carroll, 1993), and is obviously the central ability domain for understanding the measurement of reading achievement by standardized reading tests. The complete *Grw* ability taxonomy, as presented by McGrew (1997), is depicted in Figure 2. In order to evaluate and compare the results of the individual reading tests of the K-ABC, K-TEA, MBA, PIAT-R, WRAT-3, WIAT, and WJ-R/WDRB it is important to identify the extent to

which each test measures the narrow abilities of Reading Decoding (RD), Reading Comprehension (RC), and Reading Speed (RS).

It is important to note that Carroll's Reading Decoding (RD) factor subsumes both the sight reading of familiar words and the decoding of unfamiliar or nonsense words, two components of reading that may represent different constructs underneath the RD factor in Figure 2. The reader should also note that this author has taken the liberty to combine Carroll's V, RC, and CZ factors into a single Reading Comprehension (RC) factor. A careful reading of Carroll's description of these factors, logical analyses, as well as the results of analyses to be reported later in this paper, suggest that these factors may all be different "method" factors under the domain of reading comprehension.



Definition of Narrow (stratum I) *Grw* Abilities

Reading Decoding (RD) - ability to recognize and decode words or pseudowords in reading.
Verbal (printed) Language Comprehension (V) - general development, or the understanding of words, sentences, and paragraphs in native language, as measured by reading vocabulary and reading comprehension tests.
Reading Comprehension (RC) - ability to comprehend connected discourse during reading.
Cloze Ability (CZ) - ability to supply words deleted from prose passages that must be read.
Reading Speed (RS) - time required to silently read a passage as quickly as possible
Spelling Ability (SG) - ability to spell
Writing Ability (WA) - ability to write with clarity of thought, organization, and good sentence structure.
English Usage Knowledge (EU) - knowledge of writing in the English language with respect to capitalization, punctuation, usage, and spelling.

These three factors identified by Carroll may all represent a Reading Comprehension (RC) factor that is measured via three different methods (i.e., method factors)

Figure 2 - Definition of narrow (stratum I) abilities included under broad (stratum II) *Grw* ability

Three additional *Gf-Gc* abilities may also be important in an evaluation of the constructs measured by the individual reading tests covered in this review. Considerable research evidence has demonstrated the importance of phonological awareness or processing in the development of reading skills during childhood (McBride-Chang, 1995; Stahl & Murray, 1994; Torgeson, Wagner & Rashotte, 1994; Wagner & Torgeson, 1987; Wagner, Torgeson, Simmons & Rashotte, 1993).

These auditory abilities are included under the broad *Ga* domain. In addition, reading comprehension has been found to be influenced by abilities included under *Gc* (e.g., lexical knowledge, general language development) and *Gf* (e.g., ability to reason and draw inferences) (Aaron, 1995; Johnston, 1983; Lohman, 1989; McGrew & Flanagan, 1997; McGrew et al., in press; Perfetti, 1994). Although other *Gf-Gc* abilities may also be important in explaining reading performance, *Ga*, *Gc*, and *Gf* abilities are those abilities that are most likely to influence an individual's performance on a reading test at the time of testing.

MAPPING READING ACHIEVEMENT TESTS TO THE *Gf-Gc* TAXONOMY

In his book in which he outlines the Three Stratum Theory of Cognitive Abilities, Carroll (1993) comments on the broad *Gf-Gc* abilities that are most likely measured by the major intelligence tests. However, it was Woodcock's (1990) Horn-Cattell *Gf-Gc* based joint or cross-battery confirmatory factor analyses of the major individually administered intelligence test batteries that began in earnest the mapping of psychoeducational tests according to the *Gf-Gc* taxonomy.

Additional *Gf-Gc* conceptualized confirmatory cross-battery studies and logical analyses (Flanagan & McGrew, in press; McGhee, 1993; McGrew, 1997; McGrew & Flanagan, 1997) have provided useful information on the different broad and narrow cognitive abilities measured by intelligence batteries. However, little has been done in the way of joint factor analyses studies of individually administered standardized achievement test batteries.

A Review and Reanalysis of Select Achievement Battery Construct Validity Research

Studies that are useful in the examination of the *Gf-Gc* abilities measured by achievement tests must possess a number of critical characteristics. First, the achievement batteries of interest must all be administered to a common set of subjects. Second, valid indicators of other related cognitive constructs (viz., *Ga*, *Gc*, *Gf* in the case of reading) also need to be administered to the same subjects. Finally, confirmatory factor analyses methods should be used since they allow for the specification and testing of the relations between the various tests and factors (i.e., the measurement model) within the *Gf-Gc* framework.

A search of the published literature over the past five years found few studies or data sets that meet these criteria. Many studies were located that reported simple correlations between one or more achievement batteries (e.g., see Gentry, Sapp & Daw, 1995), including the concurrent validity correlations reported in the technical manuals of the seven achievement batteries covered in this review. In addition, studies (including those reported in test technical manuals) were found that reported correlations between one or more intelligence battery and a single achievement battery. Also, when studies were located that appeared promising (i.e., a sample of subjects that had been administered at least one intelligence battery that contained good *Gf-Gc* marker tests and two to three achievement batteries), the data were not subjected to any form of factor

analyses and the authors did not report a symmetrical correlation matrix (e.g., see Slate, 1996) which could be used by this author as input for independent confirmatory factor analyses of the data.

In addition to a search of the literature, this author contacted the publishers of the achievement batteries reviewed in this paper and requested copies of published or unpublished manuscripts or data (e.g., symmetrical correlation's matrices that could be independently analyzed) that meet the above study design criteria. Although the test technical manuals that were secured included much in the way of concurrent correlation information, most often the published data was not in a form that could be used. Concurrent validity correlation's between different achievement and intelligence tests, often for different samples of subjects, were almost always presented in a tabular and not symmetrical correlation format. The data could thus not be reanalyzed. Also, no useful joint factor analyses studies of more than one achievement battery were reported.

The current attempt to locate a comprehensive set of published research and data (that could be reanalyzed) that would shed light on the similarities and differences in abilities measured by different reading achievement tests was largely unsuccessful. This conclusion is similar to that of Slate (1996) who, after an "extensive literature search" (p.87), reported that he could find no published data on the interrelations between four commonly used general achievement batteries.

However, all is not lost. Five different research reports or data sets were located that meet enough of the necessary design criteria to warrant attention. Reanalysis and/or interpretation of these five data sets follows below.

The Yale Center for Learning and Attention Disorders Study

Fletcher et al. (1994) summarize the results of confirmatory factor analysis of a number of achievement batteries in a large sample of over 370 children who were part of a study conducted by the Yale Center for Learning and Attention Disorders. Although the reported analyses only included reading tests from two of the batteries currently being reviewed (viz., WJ-R/WDRB and WRAT-R) and did not include indicators of other important *Gf-Gc* abilities (nor was it conceptualized from the perspective of *Gf-Gc* theory), the results are informative, particularly as they relate to examining reading tests that use the modified cloze approach to measuring reading comprehension.

In the best fitting factor model, the WRAT-R Reading and WJ (the 1977 edition) Word Identification tests loaded very high (.97 and .95, respective), together with another word identification test, on a word identification factor (viz., Real Words factor). The magnitude of these loadings[4] suggest that the respective WRAT-R[5] and WJ Word Identification tests are very

[4] A tests *factor loading* indicates the degree of relation between the test and the latent factor. Loadings are similar to correlations and typically range from -1.0 to +1.0. The square of each loading can be used to compare loadings. For example, although factor loadings of .80 and .90 for two tests on the same factor differ by .10, the test with the .80 loading has .64% ($.80^2$ x 100) of its variance accounted for by the factor, while the test with the .90 loading has 81% ($.90^2$ x 100) of its variance accounted for by the same factor. Thus, the second test has 17% (81% - 64%) more variance attributed to the latent factor.

similar in their measurement of word identification skills. The WJ Word Attack test loaded very high (.87) with the Decoding Skills Test--Pseudowords (.92) on a decoding factor (viz., Nonsense Words), a finding that supports the construct validity of the WJ Word Attack test.

The most important finding, was the factor loading of .93 for the WJ Passage Comprehension test on a reading comprehension factor. This factor was also defined by reading comprehension measures from the Formal Reading Inventory[6] (FRI; Wiederholt, 1986) and the Gray Oral Reading Test--Revised[7] (GORT-R; Wiederholt & Bryant, 1986), and an FRI measure of listening comprehension. These three comprehension measures had factor loadings of .84, .83, and .56, respectively. The finding that the WJ Passage Comprehension test was the highest loading test (i.e., .93) on a comprehension factor for which the two other most salient factor indicators were measures of silent or oral reading comprehension that employed a different response format (viz., responding to multiple choice comprehension questions), *indicates that in this particular sample the modified cloze response format (as operationalized by the WJ/WJ-R/WDRB Passage Comprehension test) appears to measure the same comprehension construct as is measured by other reading comprehension formats.*

Although extensive cognitive data were also reported for the same sample, the respective achievement and cognitive data were not presented jointly, a condition necessary for investigating the extent to which the WRAT-R and WJ reading tests may be measuring other related cognitive constructs.

Gf-Gc Cross-Battery Reanalysis of Four Data Sets

The Samples

Data from four samples were secured that meet most of the design criteria specified above. All but the MBA and WIAT reading tests were included in at least one of these data sets.

- *Prewett and Giannuli Sample.* Prewett and Giannuli (1991) reported the results of an investigation of the relations between the reading subtests of the WJ-R/WDRB, PIAT-R, K-TEA, and WRAT-R in a sample of 118 elementary school aged subjects who had been referred for evaluation. This study did not meet all of the necessary study design criteria (i.e., it did not include indicators of other related cognitive abilities such as *Ga, Gf,* and *Gc),* but it is still an important study given that it included reading tests from four of the general achievement batteries that are being reviewed in this paper. Also, the published article reported the correlation matrix and thus, it was possible to reanalyze the data. Prewett and Giannuli used exploratory principal components

[5] Although these findings provide validity evidence for the WRAT-R Reading test, when all forms of psychometric information are considered (e.g., reliability, standardization), the WRAT-R test is judged to not meet adequate psychometric standards (see prior discussion in this paper).

[6] The FRI is an individually administered, norm-referenced measure of oral and silent reading. Two forms are read silently and two forms are read orally. Examinees are required to read passages that are followed by multiple-choice comprehension questions.

[7] The GORT-R is an individually administered, norm-referenced measure of oral reading. Examinees are required to read passages orally and then respond to a set of questions for each passage.

analysis (with an orthogonal varimax rotation) and concluded that a single large general reading component was present in the data. All the individual reading subtests (both word identification and comprehension) consistently displayed very high loadings (.88 to .96) on the general reading component.

- *WJ-R/WDRB Grade 3 Validity Sample.* This sample is described in McGrew et al. (1991) and is a nonclinical sample of 73 third grade students. In addition to indicators of the major *Gf-Gc* abilities being present, the PIAT-R and WJ-R/WDRB tests (except Reading Vocabulary) were administered to all subjects. The original data was reanalyzed.
- *WJ-R/WDRB Grade 3/4 and 10/11 Validity Samples.* Probably the most informative data sets for this paper are two concurrent validity samples reported in the WJ-R technical manual (McGrew et al., 1991). Two nonclinical samples of students in Grades 3/4 and Grades 10/11 were administered a wide array of achievement and cognitive measures. Although the size of the samples (n = 71 and 51) is of concern for multivariate analyses, in all other respects these two samples meet all the necessary design criteria for investigating the constructs measured by different reading tests. In addition to indicators of the major *Gf-Gc* abilities being present, the K-TEA, PIAT-R, WRAT-R, and WJ-R/WDRB tests were administered to both samples. In addition, the K-ABC was administered to the subjects in the Grade 3/4 sample. The original data were reanalyzed.

The Reanalysis Method and Strategy

Confirmatory factor analyses procedures were used to evaluate the extent to which reading (decoding and comprehension), fluid intelligence or reasoning (*Gf*), crystallized intelligence (*Gc*), and auditory processing (*Ga*) abilities are measured by the reading achievement tests. Briefly, in all samples (except the Prewett and Giannuli sample) a model was specified that included the latent factors (constructs) of Reading Decoding (*Grw-RD*), Reading Comprehension (*Grw-RC*), Crystallized Intelligence (*Gc*), Fluid Intelligence (*Gf*), and Auditory Processing (*Ga*). [The *Gc, Gf, Ga* factors were not present in the Prewett and Giannuli sample]. Support for the specification of the *Gf, Gc,* and *Ga* factors was presented previously. Although there is not unanimous agreement on the psychometric dimensionality of reading (Carroll, 1993; Davis, 1944, 1968, 1972; Drahozal & Hanna, 1978; Johnston, 1983; Spearitt, 1972; Thorndike, 1974), there are both practical, theoretical, and empirically supported arguments (Aaron, 1995; Gough, Hoover, Peterson, 1996; Hoover & Gough, 1990; Joshi, 1995; Johnston, 1983; Perfetti, 1994) for specifying separate RD and RC factors when attempting to identify the constructs measured by reading tests.

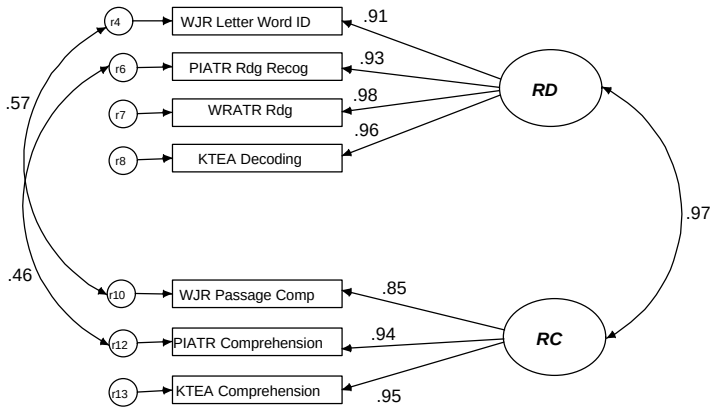
Thus, all word identification and comprehension tests were specified to load on the *Grw-RD* and *Grw-RC* factors respectively. In addition, all reading tests were initially specified to also have secondary loadings (paths) on the three cognitive factors (i.e., *Gf, Gc, Ga*). The results of each initial model were reviewed, nonsignificant paths (loadings) were deleted, new paths were added that were suggested by the program's modification indices (and that made logical and theoretical

sense), and the resultant model was re-estimated. This iterative process continued until all possible appropriate model adjustments were exhausted. It must be kept in mind that this data-driven approach to using confirmatory factor procedures is not recommended when attempting to evaluate and test theoretical models. Rather, this strategy reflects the use of confirmatory factor analysis procedures in an exploratory "model generating" mode (Jöreskog & Sörbom, 1993), a process often used in applied psychometrics for test development and evaluation.

A schematic representation of a model and the subsequent results (for the Prewett and Giannuli sample) are presented in Figure 3.[8] It is beyond the scope of this paper to explain in detail all aspects of the modeling process used, the detailed results and interpretation and, to explain in detail the use of confirmatory factor analysis procedures. The interested reader is referred to Keith's (1997) conceptual description and explanation on how to use and interpret the results of confirmatory factor analysis procedures when examining the constructs measured by psychoeducational tests.

Figure 3.

Prewett & Giannuli (1991)
Sample (n=118)
Final Model



Model Fit Statistics

Chi-square = 54.662 df = 11 p = .000
GFI = .893 AGFI = .727 PGFI = .351
CFI = .967 PCFI = .507 rmsea = .184

[8] The two-factor solution differs from the single-factor reported by Prewett and Giannuli. Appropriately designed confirmatory factor analysis research should be based on prior research and theory. See Carroll (1993) for support for the two-factor model.

Summary and Interpretation of Results

The results for each of the cross-battery factor analyses are presented in path diagram form in Appendix A. The salient factor loadings from these models, as well as the factor loadings reported in the Yale sample, are all summarized in Table 1. The results presented in Table 1, and the median (Mdn) factor loadings in particular, suggest a number of conclusions. These conclusions must be viewed with a degree of caution given the small number of samples, the limited age ranges covered, the relatively small sample sizes, and the fact that some of the average factor loadings are based on only one or two samples. The degree of confidence placed in the average factor loadings should be directly related to the number of average factor loadings upon which the average loading is based.

Table 1 - Summary of Salient Gf-Gc Factor Loadings for Reading Achievment Tests in Gf-Gc Cross-Battery Factor Analysis in Select Samples
(Analyses Completed by Kevin McGrew for NAS Committee on the Prevention of Reading Disabilities in Young Children)

	Reading Decoding (Grw-RD)							***Reading Comprehensnion (Grw-RC)***						
	Sample							***Sample***						
Test	**A**	**B**	**C**	**D**	**E**	**Mdn**	***n***	**A**	**B**	**C**	**D**	**E**	**Mdn**	***n***
K-TEA Reading Decoding		0.96		0.97	0.60	**0.96**	3		0		0	0	**0**	3
K-TEA Reading Comprehension		0		0	0	**0**	3		0.95		0.83	0.68	**0.83**	3
K-ABC Reading/Decoding				0.88		**0.88**	1				0		**0**	1
K-ABC Reading/Understanding				0		**0**	1				0.50		**0.50**	1
PIAT-R Reading Recogntion		0.93	0.96	0.92	0.92	**0.92**	4		0	0	0	0	**0**	4
PIAT-R Reading Comprehension		0	0	0	0	**0**	4		0.94	0.91	0.77	0.36	**0.84**	4
WRAT3 Reading	0.97	0.98		0.95	0.86	**0.96**	4	0	0		0	0	**0**	4
WJ-R/WDRB Letter-Word Iden.	*0.95*	0.91	0.94	0.90	0.54	**0.91**	5	0	0	0	0	0	**0**	5
WJ-R/WDRB Word Attack	*0.87*		0.84	0.76	0.56	**0.80**	4	0		0	0	0	**0**	4
WJ-R/WDRB Reading Vocab.				0	0	**0**	2	0			0.45	0.51	**0.45**	3
WJ-R/WDRB Passage Comp.	0	0	0	0	0	**0**	5	0.93	0.85	0.63	0.93	0.71	**0.85**	5

Samples
A = Yale
B = Prewett & Giannuli
C = Gr 3 WJR Val.
D = Gr 3/4 WJR Val.
E = Gr 10/11 WJR Val.

Given these caveats, based on the average factor loadings presented in Table 1, the following conclusions are offered:

- The K-TEA Reading Decoding, K-ABC Reading Decoding, PIAT-R Reading Recognition, WRAT-R(3)[9], and WJ-R/WDRB Letter-Word Identification tests appear to be equally valid indicators of the Reading Decoding construct. Performance on these five word identification tests was also not found to be significantly influenced (in any sample) by the abilities of *Gf, Gc, or Ga*.

[9] See footnote 1.

- The WJ-R/WDRB Word Attack test is also an empirically strong indicator of Reading Decoding (RD), although it had the lowest average RD factor loading (.80). This most likely reflects the fact that the WJ-R/WDRB Word Attack test measures a different subcomponent of RD than the other five word identification tests (viz., the ability to apply phonic and structural analysis skills to words).
- Despite using different response formats to measure reading comprehension (viz., open-ended verbal/gestural; multiple choice; modified cloze), the K-TEA Reading Comprehension, PIAT-R Reading Comprehension, and WJ-R/WDRB Passage Comprehension tests appear to all be equally strong and valid (median factor loadings of .83, .84, and .85, respectively) indicators of Reading Comprehension (RC). None of these reading comprehension tests displayed significant average (median) factor loadings on the *Gf*, *Gc*, and *Ga* factors.
    - The K-ABC Reading Understanding test may measure RC differently than the other comprehension tests. It displayed a factor loading of .50 in a sample (D) where the other respective comprehension tests displayed factor loadings from .77 to .93. The K-ABC Reading Understanding test also appears to be significantly influenced by Crystallized Intelligence (*Gc* loading of .32). Given that these findings are based on only one sample, this conclusion must be viewed cautiously. Replication in additional samples is needed.
- The WJ-R/WDRB Reading Vocabulary test appears to be the least "pure" measure of reading. It appears to be a mixed measure of both RC (Mdn = .45), *Gc* (Mdn = .38), and *Gf* (Mdn = .10). Although these average values are based on a small number of samples (2-3), the results make logical sense as the Reading Vocabulary test requires an examinee to read and comprehend the meaning of words (*Grw-RD* and *Gc*), and then reason by providing an antonym or synonym (*Gf* and *Gc*).
- Inspection of the correlation between the Reading Decoding (RD) and Reading Comprehension (RC) latent factors and the respective loadings of RD and RC on a second-order broad reading factor (results not presented in Table 1 - see path diagrams in Appendix A and B), indicates that RD and RC are highly correlated factors/constructs. This does not necessary mean they are identical constructs, as factor or structural evidence is only one form of evidence for cognitive constructs. Other forms of evidence (i.e., developmental, predictive, neuropsychological, and heritability) (Horn & Noll, 1997) are needed to form a complete network of construct validity evidence. The identification and separation of the RD and RC factors in Carroll's (1993) analyses support this distinction.
- Although not the focus of this paper, the findings reported for *Ga* and RD warrant comment. As discussed previously, considerable research has demonstrated the importance of phonological awareness or processing (abilities subsumed by *Ga*) in the development of reading skills during childhood (McBride-Chang, 1995; Stahl & Murray, 1994; Torgeson et al., 1994; Wagner & Torgeson, 1987; Wagner et al., 1993). The significant loadings of the WJ-R Word Attack test on the *Ga* factor in the norm samples (see Table 2) are consistent with this research. The lack of such a relation in the

elementary school-age special samples (see Table 1), plus the significant WJ-R Letter-Word Identification/*Ga* loading in the Grade 10/11 special sample (see Table 1), are not consistent with this research. A number of possible hypothesis exist for these findings:

- Given the lack of a similar finding in the literature with the same WJ-R tests (see McGrew, 1993 and McGrew et al., in press), the significant Grade 10/11 WJ-R Letter-Word Identification/*Ga* path most likely is a chance finding due to sampling error.
- The relation between *Ga* and reading is most strong during the early elementary school years and has been shown to decrease in importance with age when investigated with the WJ-R tests (see McGrew, 1993 and McGrew et al., in press). The significant Word Attack/*Ga* findings in the grade 1-3 WJ-R norm samples (see Table 2) are consistent with this finding. The nonsignificant results for the special samples (see Table 1) were primarily in small samples of students in grades 3 and above. Also, concerns have previously been expressed about the representativeness of two of the samples by the authors (see McGrew et al., 1991). The results based on the large nationally representative WJ-R norm samples are those that are most likely to cross-validate.
- A review of the path diagrams in Appendix B reveals a potentially important difference between the current analyses and other studies that have investigated the relations between measures of phonological awareness and reading. Much of the existing reading research has not investigated the importance of phonological awareness within the context of a comprehensive hierarchical model of intelligence (e.g., contemporary *Gf-Gc* theory). In the path diagrams in Appendix B (diagrams that do not include the complete *Gf-Gc* model), Word Attack (the most likely reading skill influenced by phonological awareness) is explained by a combination of direct *and* indirect influences. Direct influences are in the form of direct paths from the *Ga* and RD factors to the Word Attack test. Indirect effects are also present through the influence of *g* as mediated through, *Ga*, *Gr*, and RD. Most reading research has only included direct effects (often without the specification of direct effects of higher-order abilities within the domain of reading), and infrequently has included indirect effects of *g*, and other *Gf-Gc* abilities. Analyses of the same WJ-R norm data with regression methods (with only direct effects present in the models) (McGrew, 1993) and with structural equation modeling methods (that included both hierarchical intelligence and reading achievement models with direct and indirect effects in a single causal model) (McGrew et al., in press) produced different findings, with the relative importance of specific (e.g., *Ga*) and general (*g*) cognitive abilities varying as a function of age and the breadth of the reading achievement construct predicted/explained. Thus, an integration of the results reported for *Ga* abilities in this paper with other research must be done with an understanding of important differences in samples and research methodology.

THE CLOZE APPROACH TO MEASURING READING COMPREHENSION: IS IT POSSIBLE TO "CLOZE" THE BOOK ON QUESTIONS THAT HAVE BEEN RAISED ABOUT THIS PROCEDURE?

Definition

The *cloze approach* or procedure can be defined as "a method of systematically deleting words from a prose selection and then evaluating the success a reader has in accurately supplying the words deleted" (McKenna & Robinson, 1980). The assumption underlying the cloze procedure is that a reader can only supply the correct word if he or she understands (i.e., comprehends) the meaning of the text (Joshi, 1995). The cloze procedure is an attempt to "assess reading comprehension by providing longer, hence more 'real world,' reading experiences during assessment than is possible with other formats" (Osterlind, 1989).

Although a form of the cloze technique was used as long ago as 1897 to investigate memory, Taylor (1953) is credited with first developing and applying this procedure to the measurement of language or reading proficiency (Carroll, 1993; McKenna & Robinson, 1980). Variants of the cloze procedure have been used for a variety of applications (e.g., foreign language; determining readability of text; a teaching device) (McKenna & Robinson, 1980). It is the application of the cloze procedure to the measurement of reading comprehension that is the current focus. Readers interested in an in-depth knowledge of the cloze approach should consult McKenna and Robinson's (1980) annotated bibliography that covers (a) background, (b) literature reviews, (c) comprehension and readability, (d) statistical and constructional issues, (e) the psychology of cloze, (f) contextual phenomena, (g) use as a teaching device, (h), foreign language applications, (i) and the cloze and maze procedure.

The Question

No approach to the measurement of the construct of reading is without its advantages and disadvantages (Johnston, 1983; Joshi, 1985) The purpose here is not to review in detail the various pros and cons that have been listed for the cloze approach, but to address the most frequently asked question about this approach. The issue that is most frequently raised about the cloze approach is whether the "technique measures reading or language comprehension in the same way as more conventional tests" (Carroll, 1993, p. 167). A more specific question that is often asked is whether cloze tests measure literal or inferential comprehension (Woodcock, 1997). Critics of the cloze approach often state that although it may adequately measure the comprehension of sentences, "it cannot be used to test comprehension of connected passages and cohesive texts" (Joshi, 1995, p.366).

A Literature Review-Based Answer

Carroll (1993) reports that there is little *factorial* literature available that can shed light on whether the cloze approach measures the same aspects of reading comprehension as measured by

other "traditional" formats. In one of the few factor-based studies, Bormuth (1969) reported the finding of a large predominant trait in a cloze measure that was interpreted to represent reading comprehension. Bachman (1982) reported a confirmatory factor analysis study of cloze passages that used a "rational" (in contrast to the more traditional "random") word deletion strategy in the construction of the passages. Bachman (1982) concluded that "cloze tests can be used to measure higher order skills--cohesion and coherence--if a rational deletion procedure is followed....cloze passages using a rational deletion procedure can be used to measure textual relationships" (p. 67). However, Carroll (1993) reanalyzed Bachman's data (with an exploratory factor approach) and was not in complete agreement with Bachman's conclusions. Carroll (1993) reported that he found the same number of factors as Bachman, but Carroll reported that his factors were not as clearly interpretable.

Although factor-based research may be limited, early on in the reading literature there were a significant number of concurrent validity studies. Bormuth (1962, 1963) found very strong and significant correlation's (e.g., r = .96) between cloze and a multiple-choice measures of reading comprehension, a finding that Rankin and Culhane (1969) later replicated. Other studies have also reported significant concurrent validity correlation's with measures of various types of comprehension (e.g., main idea, inferential and literal comprehension, literary awareness) (Gallant, 1965; Rankin, 1958; Smith & Zinc, 1977), as well as with scores generated from the miscue analysis of oral reading (Page, 1977). Finally, in their annotated bibliography, McKenna and Robinson (1983) conclude that the available research suggests that "rather convincingly...cloze does, in fact, measure comprehension" (p.12).

A Logically-Based Answer

An explanation that might account for some of the divergent opinions regarding the cloze approach is that there are differences between different cloze procedures that may impact the ability of the different procedures to validly measure reading comprehension. That is, "all cloze tests may not be created equal." A significant problem is the fact that no set rule or criteria have been established for best selecting words for removal in the sentences, phrases, or passages (Osterlind, 1989). The traditional procedure calls for the deletion of every *n*th (e.g., fifth, seventh) word in the text. The defenders of the cloze procedure acknowledge that traditional cloze measures may indeed be less than optimal measures of reading comprehension and, that a more rational deletion strategy (Bachman, 1982) that takes into consideration the "syntactic, semantic, and graphonic context at levels from word to discourse" (Johnston, 1983, p. 63) can measure reading comprehension very well.

The most prominent contemporary proponent of the *modified cloze* procedure, as reflected in his consistent use of this procedure as the primary indicator of reading comprehension in his family of test batteries that include measures of reading (viz., the MBA; WRMT-R; WDRB; WJ-R) (Woodcock, et al., 1994; Woodcock, 1987; Woodcock, 1997; Woodcock & Johnson, 1989), is Richard Woodcock. Woodcock's position (Hessler, 1993; Woodcock, 1997) is that a modified cloze test (viz., one constructed through the careful selection of words to delete so that the correct response requires comprehension of the entire passage and, where the subject is unable to

provide the answer based only on local context of the passage) is a valid approach to the measurement of reading comprehension. Woodcock (1997, p. 49) presents the following example to demonstrate the content validity of the modified cloze procedure.

> "Try responding to the following three cloze examples.
>
> '...do something about _____ it.'
>
> Now, read the entire sentence and attempt to provide the missing word:
>
> 'It is another to do something about _____ it.'
>
> Finally, read and answer the entire item that is comprised of two sentences:
>
> 'It is one thing to demonstrate that modern war is harmful to the species. It is another thing to do something about _____ it.'
>
> Note that the solution to this item (for example, the word *preventing*) required understanding not only of the sentence containing the blank, but also the preceding sentence."

Confirmatory Factor-Based Answer # 1

If the modified cloze procedure outlined and reflected in the WJ-R/WDRB Passage Comprehension test is indeed a valid measure of reading comprehension as argued by Woodcock, then it would be predicted that this test's loading on a reading latent trait or factor would be of similar in magnitude to the factor loadings of other reading comprehension test indicators (e.g., PIAT-R Reading Comprehension) that also define the comprehension factor. Based on the results summarized in Table 1 and discussed previously, the available cross-battery factor analyses research confirms this prediction. The modified cloze procedure (as reflected in the WJ-R/WDRB Passage Comprehension test) was found to have an average Reading Comprehension (RC) factor loading of .85, a value very similar to the average loadings of .83 and .84 for the K-TEA and PIAT-R reading comprehension tests, tests that use different formats. *Gf-Gc conceptualized cross-battery confirmatory factor analysis studies support the construct validity of the modified cloze procedure as an indicator of reading comprehension.*

Confirmatory Factor-Based Answer # 2

A final series of special confirmatory factor analyses were completed to further examine the construct validity of the modified cloze procedure. Models that included *Gf*, *Gc*, *Ga*, *Grw-RD*, and *Grw-RC* factors were evaluated in four large (n = 157 to 351) nationally representative samples (the WJ-R/WDRB norm data) from kindergarten through grade four (to reflect the NAS committee focus on the prevention of reading disabilities in young children). Since other non-WJ-R/WDRB reading test indicators were not present in the data, the models can not be used to evaluate the construct validity of the modified cloze procedure with other reading comprehension indicators. The value of these analyses is the ability to determine if the WJ-R/WDRB Passage Comprehension test (as well as the other three WJ-R/WDRB reading tests) is measuring other construct irrelevant variance (i.e., variance not due to reading). The path diagrams for the final models are included in Appendix B. The results of these analyses are summarized in Table 2.

Inspection of the average (Mdn) factor loadings in Table 2 suggests the following conclusions:

- The WJ-R/WDRB Passage Comprehension test does not appear to measure construct irrelevant *Gc*, *Gf*, or *Ga* variance.
- The WJ-R/WDRB Letter-Word Identification test does not appear to measure construct irrelevant *Gc*, *Gf*, or *Ga* variance.
- In addition to being a measure of Reading Decoding (RD), the WJ-R/WDRB Word Attack test may also measure *Ga* abilities (Mdn loading of .16)
- In addition to being a measure of Reading Comprehension (RC), the WJ-R/WDRB Reading Vocabulary test may also measure a small amount of *Gc* (Mdn loading of .08). *Gc* abilities appear to increase in importance on Reading Vocabulary with increasing age (i.e., nonsignificant loadings in kindergarten and grade one; loadings of .16 and .23 in second and third grades, respectively).

Table 2.

	Crystallized Intelligence (Gc)					***Fluid Intelligence (Gf)***					***Auditory Processing (Ga)***				
	Samples					***Samples***					***Samples***				
Test	**C**	**D**	**E**	**Mdn**	***n***	**C**	**D**	**E**	**Mdn**	***n***	**C**	**D**	**E**	**Mdn**	***n***
K-TEA Reading Decoding		0	0	**0**	2		0	0	**0**	2		0	0	**0**	2
K-TEA Reading Comprehension		0	0	**0**	2		0	0	**0**	2		0	0	**0**	2
K-ABC Reading/Decoding		0		**0**	1		0		**0**	1		0		**0**	1
K-ABC Reading/Understanding		0.32		**0.32**	1		0		**0**	1		0		**0**	1
PIAT-R Reading Recogntion	0	0	0	**0**	3	0	0	0	**0**	3	0	0	0	**0**	3
PIAT-R Reading Comprehension	0	0	0.43	**0**	3	0	0	0	**0**	3	0	0	0	**0**	3
WRAT3 Reading		0	0	**0**	2		0	0	**0**	2		0	0	**0**	3
WJ-R/WDRB Letter-Word Iden.	0	0	0	**0**	3	0	0	0	**0**	3	0	0	0.25	**0**	3
WJ-R/WDRB Word Attack	0	0	0	**0**	3	0	0	0	**0**	3	0	0	0	**0**	3
WJ-R/WDRB Reading Vocab.		0.35	0.42	**0.38**	2		0.19	0	**0.10**	2		0	0	**0**	2
WJ-R/WDRB Passage Comp.	0	0	0	**0**	3	0.23	0	0	**0**	3	0.18	0	0	**0**	3

NOTES: *n= number of samples upon which the average (median; Mdn) factor loadings are based. Two loadings in italics for WJ-R/WDRB Lettter-Word Iden* & Word *Attack tests indicate that each test loaded on separate reading decoding factors in Sample A ("nonsense" and "read" words). Loadings of 0 indicate that the test had a nonsignificant loading (prescribed to be "0" in the confirmatory model - i.e., no path) in the respective analyses.*

An Integrated Answer

The primary criticism and question that is directed at the modified cloze procedure, which is often based on face validity analyses, is not supported by the available research literature. The evidence reviewed here suggests that the modified cloze approach (as reflected by the WJ-R/WDRB Passage Comprehension test) is a valid approach to the measurement of reading comprehension. Stated differently, the WJ-R/WDRB Passage Comprehension test is just as good of an indicator of reading comprehension as are other reading tests that use other formats (viz.,

K-TEA, K-ABC, and PIAT-R). The WJ-R/WDRB Passage Comprehension test shows strong average loadings on a multiple-indicator based Reading Comprehension (RC) factor and does not appear to measure extraneous construct irrelevant variance (viz., *Gf, Gc,* and *Ga).*[10]

SUMMARY AND CONCLUSIONS

The measurement of hypothetical mental constructs (e.g., reading) is an imperfect process. Conclusions about unobservable latent traits (e.g., reading) can only be inferred indirectly from samples of observable behavior. In general, the greater the number and range of behavior samples secured the more confidence one can place in the accuracy and validity of the inferences about the construct of interest. This is particularly true about broad constructs (e.g., reading) that are conceptualized as being comprised of a number of different, yet correlated components. The optimal operationalization of the construct of reading would be the use of multiple, reliable, nonredundant indicators of the subcomponents that comprise reading (e.g., word identification, word attack, reading speed, and comprehension).

A composite score based on the combination of all the individual reading tests from all batteries reviewed in this paper would represent a very reliable and valid (but not efficient) indicator of the construct of reading. However, this is not a practical option. As a result, test developers have attempted to develop efficient reading tests that still meet acceptable standards of reliability and validity. In the process, different reading tests have been constructed that have much in common (i.e., shared construct relevant variance), but that also differ in how much they share with other cognitive constructs (i.e., construct irrelevant variance) and how much they measure that is unique (i.e., unique variance). Thus, the process of determining if different reading tests are measuring similar constructs (i.e., are we comparing apples to apples, or apples to oranges ?), is largely a process of partitioning and comparing reading test score variance.

The Construct Validity of Reading Tests: Apples, not Oranges

The cross-battery factor analyses reported in this paper indicate that most of the individual reading tests reviewed are indeed valid indicators of the construct of reading (i.e., they are all apples).[11] The median factor loadings for the different tests on the reading decoding and comprehension factors were, for the most part, consistently in the .80s to .90s. Although these values are very high, when properly interpreted, these data indicate that the different reading tests, although all still apples, may represent different types of apples.

In the case of comprehension, with the exception of the K-ABC Reading Understanding and WJ-R/WDRB Reading Vocabulary tests, the median factor loadings for the respective

[10] There are differing opinions regarding whether the presence or absence of construct irrelevant variance in reading comprehension tests is "good" or "bad". Although reading comprehension tests with little construct irrelevant variance may provide for the purest measurement of the construct of reading, tests that draw upon other abilities (e.g., *Gf, Gc*)in order to answer inferential questions may have equally important real-world predictive value. Both types of tests may be useful for different purposes.

[11] The reader is asked to grant this author a few "degrees of freedom" to continue the use of the apples-to-oranges analogy. The intent is to ground psychometric terms and concepts in a concrete example.

comprehension tests on the Reading Comprehension (RC) factor ranged from .83 to .85. The square of these factor loadings represents the proportion or percent (68% to 72%) of common or shared RC variance. Common factor variance this high is strong evidence for the construct validity of the respective reading comprehension tests. The *K-TEA Reading Comprehension, PIAT-R Reading Comprehension, and WJ-R/WDRB Passage Comprehension tests are all valid indicators of the construct of reading comprehension*. They are all equally valid indicators of the same variety of fruit (i.e., reading comprehension).

However, this does not mean that all of the reading comprehension tests are identical. Although approximately 70% shared variance for the respective reading comprehension tests is high, there still remains approximately 30% variance that each reading comprehension test does not share in common with the other reading comprehension tests. A portion of this remaining variance reflects the unreliability or error variance of each test. Just as different individual red delicious apples are not perfect indicators of the variety of red delicious apples (i.e., they differ in shape, color, flaws, etc.), so do individual tests contain a degree of error.

For discussion purposes, a reasonable assumption is that the average error variance for the reading tests is 10% (1 - reliability coefficient of .90). This then suggests that the reading comprehension tests still have 20% variance that is unique to each test. It is likely that the variability in different item formats described previously accounts for a significant portion of this remaining unique variance. Just as both a red and a golden delicious apple are valid indicators of the variety of delicious apples (high shared variance) and, just as both have flaws that make them less than perfect exemplars of this variety (error variance), they are both unique in color (unique variance).

The majority of the median factor loadings for the respective word identification and word attack tests on the Reading Decoding (RD) factor ranged from .88 to .96. The square of these factor loadings indicates that the different word identification and word attack tests have approximately 80% to 90% shared variance. This high degree of communality is likely related to the very similar measurement approaches used to measure this construct across most of the different reading tests. *The respective K-TEA, K-ABC, PIAT-R, WIAT, WRAT-3*[12]*, and WJ-R/WDRB word identification and word attack tests all appear to all be strong and equally valid indicators of RD.* RD appears to be a reading construct for which it is relatively easy to design reliable and valid indicators.

Communality does not Equal Identity: A Caveat

There is little doubt that tests of word identification/attack and reading comprehension skills, within their respective domains, have much in common (i.e., high shared variance). However, this does not mean that the scores produced by tests that have empirically demonstrated construct validity will always produce similar level of performance scores. For example, studies that have included the PIAT-R and WJ-R/WDRB (Daub & Colarusso, 1996), K-TEA and WRAT-R (Bell,

[12] See footnote 1.

Lentz, & Graden, 1992), WIAT and PIAT-R (Slate, 1996), and other achievement tests not included in this review, frequently report significant mean score differences between reading scores within the same domain (e.g., reading comprehension).

Significant differences between the scores derived from reliable and construct valid tests within the same domain of reading (e.g., reading comprehension) can occur for a number of reasons. First, as described above, unique and error variance characteristics may influence different tests in different ways. Second, significant differences in the recency of the norm samples of tests may account for score differences. Tests whose norms are based on more recent standardization samples typically produce lower (but more representative) scores than tests that may have older and possibly outdated norms. Third, test score differences may reflect differences in the construction of norm samples for different tests. Fourth, differences in achievement test and curriculum content overlap (i.e., curriculum content validity) can result in a bias in either direction for different tests (Bell et al., 1992).

Thus, even though the different reading tests included in this review generally demonstrated strong evidence of being valid indicators of either reading decoding or reading comprehension, it is still possible for the scores provided by the respective tests to differ for individuals or groups of individuals. Consumers of reading tests need to recognize that, although testing technology is a well developed applied science that can produce psychometricaly sound tests of reading, it is not a perfect science. Different reliable and valid test indicators of the construct of reading cannot be expected to produce identical results. In general, most of the time the test score results will be similar, but an expectation of identity is not realistic.

The Final Analyses: Good Apples, Bad Apples, and Applesauce

A critical review of the technical adequacy and construct validity of individually administered standardized achievement batteries produces the following conclusions:

- No individual test or combinations of tests within a general achievement battery can be considered to be the best operational measure of the construct of reading.
- The K-TEA Reading Decoding, K-ABC Reading Decoding, PIAT-R Reading Recognition, and WJ-R/WDRB Letter-Word Identification tests are technically adequate and equally valid indicators of the construct of Reading Decoding. The various word identification tests were not found to contain any significant construct irrelevant variance.
- The K-TEA Reading Comprehension, PIAT-R Reading Comprehension, and WJ-R/WDRB Passage Comprehension tests are technically adequate and equally valid indicators of the construct of Reading Comprehension. These tests were not found to contain any significant construct irrelevant variance.
- The modified cloze approach to the measurement of reading comprehension, as operationalized in the WJ-R/WDRB Passage Comprehension test, is as equally a valid approach to the measurement to reading comprehension as are other approaches that use different formats (e.g., open-ended or multiple choice responses to questions).

- The WRAT-3 Reading test, although showing strong factor loadings on a reading decoding factor, is not recommended for use in applied or research settings. The instrument, as well as its predecessors, has questionable technical characteristics and measures a vary narrow aspect of reading.
- Two of the reading comprehension tests appear to be technically sound, but factorially complex measures, of both reading and crystallized intelligence (*Gc*). The WJ-R/WDRB Reading Vocabulary and K-ABC Reading Understanding tests appear to be hybrid (apple+orange?) measures of reading comprehension and crystallized intelligence.
- It was not possible to investigate the construct validity of the WIAT reading tests since no data sets could be located that would provide useful data for this paper. Based on task analysis of the Basic Reading and Reading Comprehension test items and response formats, it is predicted each of these tests would be found to be technically sound and equally valid measures of the constructs of reading decoding and comprehension, respectively. Additional research is needed to confirm this prediction.
- The MBA was included only indirectly in the factor analyses reported here. Since the MBA is comprised of three parts that are essentially mini-versions of the WJ-R/WDRB Letter-Word Identification, Reading Vocabulary, and Passage Comprehension tests, the findings for the three WJ-R/WDRB tests most likely generalize to the MBA. Given that the MBA only provides a single composite score that is based on a combination of valid indicators of reading decoding and reading comprehension, and a hybrid measure of reading comprehension and to a limited extent crystallized intelligence (*Gc*), this test can be considered to be a single indicator of the broad construct of reading. Carrying the fruit analogy one final step further, the MBA is like an applesauce that is made from a mixture of different varieties of apples (decoding and comprehension), with a touch of orange (*Gc)* flavoring.[13]

[13] The same can also be said about the K-TEA Brief test, an abbreviated version of the complete K-TEA that was included in this review.

REFERENCES

Aaron, P.G. (1995). Differential diagnosis of reading disabilities. School Psychology Review, 24(3), 345-360.

Allinder, R.M. & Fuchs, L.S. (1992). Screening academic achievement: Review of the Peabody Individual Achievement Test-Revised. Learning Disabilities Research & Practice, 7, 45-47.

Anastasi, A. (1985). Review of the Kaufman Assessment Battery for Children. In J. V. Mitchell (Ed.), The ninth mental measurements yearbook (pp. 769-771). Lincoln NE: Buros Institute.

Bachman, L.F. (1982). The trait structure of cloze test scores. TESOL Quarterly, 16, 61-70.

Barnett, D.W., & Zins, J.E. (1983). The Kaufman Assessment Battery for Children and school achievement: A validity study. Journal of Psychoeducational Assessment, 1, 235-241.

Bell, P.F., Lentz, F.E., & Graden, J.L. (1992). Effects of curriculum-test overlap on standardized achievement test scores: Identifying systematic confounds in educational decision making. School Psychology Review, 21(4), 644-655.

Bennes, K.M. (1992). Review of the Peabody Individual Achievement Test-Revised. In J.C. Conoley, & J.J. Kramer (Eds.), The eleventh mental measurements yearbook (pp. 649-652). Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln.

Bormuth, J.R. (1962). Cloze tests as measures of readability and comprehension ability. Unpublished doctoral dissertation, Indiana University

Bormuth, J.R. (1963). Cloze as a measure of readability. In J. Allen Figurel (Ed.), Reading as an intellectual activity, Proceedings of the International Reading Association, 8. New York: Scholastic Magazines.

Bormuth, J. R. (1969). Factor validity of cloze tests as measures of reading comprehension ability. Reading Research Quarterly 4, 358-365.

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

Carroll, J.B. (1997). The three-stratum theory of cognitive abilities. In D.P. Flanagan, J.L. Genshaft, & P.L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues.(pp. 122-130). New York: Guilford Press.

Clark, E. (1992). Review of the Wide Range Achievement Test--Revised. In J. J. Kramer & J. C. Conoley (Eds.), The tenth mental measurements yearbook (pp. 897-903). Lincoln NE: Buros Institute.

Coffman, W. E. (1985). Review of the Kaufman Assessment Battery for Children. In J. V. Mitchell (Ed.), The ninth mental measurements yearbook (pp. 771-773). Lincoln NE: Buros Institute.

Cooter, Jr., R.B. (1989). Review of the Woodcock Reading Mastery Tests-Revised. In J.C. Conoley, & J.J. Kramer (Eds.), The tenth mental measurements yearbook (pp. 910-913). Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln.

Cruse, C., Dumont, R., & Willis, J. (1996).. Test review: Woodcock-McGrew-Werder Mini-Battery of Achievement. NASP Communique, Sept., 22.

Daniel, M. H. (1997). Achievement test normative updates due for the 1997-98 school years. Assessment Information Exchange, 10, 1-7.

Daub, D. & Calarusso, R. P. (1996). The validity of the WJ-R, PIAT-R, and DAB-2 reading subtests with students with reading disabilities. Learning Disabilities Research and Practice, 11, 90-95.

Davis, F.B. (1944). Fundamental factors of comprehension of reading. Psychometrika, 9, 185-197.

Davis, F.B. (1968). Research in comprehension in reading. Psychometrika, 3, 499-545.

Davis, F.B.(1972). Psychometric research on comprehension in reading. Reading Research Quarterly, 7, 628-678.

Doll, E. J. (1989). Review of the Kaufman Test of Educational Achievement. In J. J. Kramer & J. C. Conoley (Eds.), The tenth mental measurements yearbook (pp. 410-412). Lincoln NE: Buros Institute.

Drahozal, E.C., & Hanna, G.S. (1978). Reading comprehension subscores: Pretty bottles for ordinary wine. Journal of Reading, 21, 416-420.

Flanagan, D.P., Genshaft, J.L., & Harrison, P.L. (Eds.) (1997). Contemporary intellectual assessment: Theories, tests and issues. New York: Guilford.

Flanagan, D.P., & McGrew, K. S. (1996). A "cross-battery" approach to assessing and interpreting cognitive abilities: An alternative to the Wechsler tradition. Manuscript submitted for publication.

Flanagan, D.P., & McGrew, K.S. (1997). A cross-battery approach to assessing and interpreting cognitive abilities: Narrowing the gap between practice and science. In D.P. Flanagan, J.L. Genshaft, & P.L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests and issues (pp. 314-325). New York: Guilford Press.

Flanagan, D.P., & McGrew, K.S. (in press). Interpreting intelligence tests from contemporary Gf-Gc theory: Joint confirmatory factor analyses of the WJ-R and KAIT in a non-white sample. Journal of School Psychology.

Fletcher, J. M., Francis, D. J., Stuebing, K. K., Shaywitz, B. A., Shaywitz, S. E., Shankweiler, D. P., Katz, L., & Morris, R. D. (1994). Conceptual and methodological issues in construct definition. In G. R. Lyon (Ed.), Frames of reference for the assessment of learning disabilities: New views on measurement issues. Baltimore: Paul H. Brookes.

Gallant, R. (1965). Use of cloze tests as a measure of readability in the primary grades. In J. Allen Figurel (Ed.) Reading and Inquiry, Proceedings of the International Reading Association, 10. Newark, Delaware: Interantional Reading Association.

Gentry, N., Sapp, G. L., & Daw, J. L. (1995). Scores on the Wechsler Individual Achievement Test and the Kaufman Test of Educational Achievement-Comprehensive Form for emotionally conflicted adolescents. Psychological Reports, 76, 607-610.

Gough, P.B., Hoover, W.A., & Peterson, C.L. (1996). Some observations on a simple view of reading. In C. Cornoldi, & J. Oakhill (Eds.), Reading comprehension difficulties: Processes and intervention. Mahwak, NJ: Lawrence Erlbaum Associates.

Gregory, R.J. (1996). Psychological testing: History, principles, and applications. Boston: Allyn & Bacon.

Gustafsson, J.-E. (1984). A unifying model for the structure of intellectual abilities. Intelligence, 8, 179-203.

Gustafsson, J.-E. (1988). Hierarchical models of individual differences in cognitive abilities. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 4, pp. 35-71). Hillsdale, NJ: Erlbaum.

Harrison, P. L. (1989). Review of the Wide Range Achievement Test--Revised. In J. J. Kramer & J. C. Conoley (Eds.), The tenth mental measurements yearbook (pp. 903-905). Lincoln NE: Buros Institute.

Harrison, P. L., Kaufman, A. S., Hickman, J. A., & Kaufman, N.L. (1988). A survey of tests used for adult assessment. Journal of Psychoeducational Assessment, 6, 188-198.

Hessler, G.L. (1993). Use and interpretation of the Woodcock-Johnson Psychoeducational Battery-Revised. Chicago: Riverside.

Hoover, W.A., & Gough, P.B. (1990). The simple view of reading. Reading and Writing: An Interdisciplinary Journal, 2, 127-160.

Horn, J. L. (1988). Thinking about human abilities. In J. R. Nesselroade & R. B. Cattell (Eds.), Handbook of multivariate psychology (rev. ed., pp. 645-865). New York: Academic Press.

Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K.S. McGrew, J.K. Werder, and R.W. Woodcock, Woodcock-Johnson Technical Manual (pp. 197-232). Chicago: Riverside.

Horn, J. L. (1994). Theory of fluid and crystallized intelligence. In R. J. Sternberg (editor in chief), Encyclopedia of human intelligence (pp. 443-451). New York: Macmillan.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In Flanagan, D.P., Genshaft, J.L., & Harrison, P.L. (Eds.), Contemporary intellectual assessment: Theories, tests and issues (pp. 53-91). New York: Guilford.

Jaeger, R.M. (1989). Review of the Woodcock Reading Mastery Tests-Revised. In J.C. Conoley, & J.J. Kramer (Eds.), The tenth mental measurements yearbook (pp. 913-916). Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln.

Johnston, P.H. (1983). Reading and comprehension assessment: A cognitive basis. Delaware: International Reading Association.

Joshi, R.M.(1995). Assessing reading and spelling skills. School Psychology Review, 24(3), 361-375.

Jöreskog, K., & Sörbom, D. (1993). LISREL 8: Structural equation modeling with the SIMPLIS command language. Hillsdale, NJ: Lawrence Erlbaum.

Kaufman, A. S. (1985). Review of the Woodcock-Johnson Psychoeducational Battery. In J. V. Mitchell (Ed.), The ninth mental measurements yearbook (pp. 1762-1765). Lincoln NE: Buros Institute.

Kaufman, A.S., & Kaufman, N.L. (1983). Kaufman Assessment Battery For Children. Circle Pines, MN: American Guidance Service.

Kaufman, A.S., & Kaufman, N.L. (1985). Kaufman Tests of Educational Achievement. Circle Pines, MN: American Guidance Service.

Keith, T. Z. (1997). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In Flanagan, D.P., Genshaft, J.L., & Harrison, P.L.

(Eds.), Contemporary intellectual assessment: Theories, tests and issues (pp. 373-402). New York: Guilford.

Lewandowski, L. J. (1986). Test review: Kaufman Test of Educational Achievement. Journal of Reading, 30, 258-261.

Lohman, D.F. (1989). Human Intelligence: An introduction to advances in theory and research. Review of Educational Research, 59(4), 333-373.

Luther, J. B. (1992). Review of the Peabody Individual Achievement Test--Revised. Journal of School Psychology, 30, 31-39.

Mabry, L. (1995). Review of the Wide Range Achievement Test 3. In J.C. Conoley & J.C. Impara (Eds.), The twelfth mental measurements yearbook (pp. 1108-1110). Lincoln, NE: Buros Institute.

Mardell-Czudnowski, C. (1996). A survey of assessment professionals in the US: Testing children with special needs. School Psychology International, 17, 189-203.

Markwardt, F.C. (1989). Peabody Individual Achievement Test-Revised. Circle Pines, MN: American Guidance Service.

McBride-Chang, C. (1995). What is phonological awareness? Journal of Educational Psychology, 87(2), 179-192.

McGhee, R. (1993). Fluid and crystallized intelligence: Confirmatory factor analysis of the Differential Abilities Scale, Detroit Tests of Learning Aptitude-3, and Woodcock-Johnson Psycho-Educational Battery - Revised. Journal of Psychoeducational Assessment Monograph Series: WJ-R Monograph, 20-38.

McGrew, K. S. (1993). The relationship between the WJ-R Gf-Gc cognitive clusters and reading achievement across the lifespan. Journal of Psychoeducational Assessment, Monograph Series: WJ-R Monograph, 39-53.

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In Flanagan, D.P., Genshaft, J.L., & Harrison, P.L. (Eds.), Contemporary intellectual assessment: Theories, tests and issues (pp. 151-179). New York: Guilford.

McGrew, K. S., & Flanagan, D. P. (1997). The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment. Boston: Allyn & Bacon.

McGrew, K. S., Flanagan, D. P., Keith, T. K., & Vanderwood, M. (in press). Beyond g: The impact of Gf-Gc specific cognitive abilities research on the future use and interpretation of intelligence test batteries in the schools. School Psychology Review.

McGrew, K. S., Werder, J. K., & Woodcock, R. W. (1991). WJ-R technical manual. Chicago: Riverside.

McKenna, M C. & Robinson (1980). An introduction to the cloze procedure: An annotated bibliography. Newark, DE: International Reading Association.

Messick, S. (1992). Multiple intelligences or multilevel intelligence? Selective emphasis on distinctive properties of hierarchy: On Gardner's Frames of Mind and Sternberg's Beyond IQ in the context of theory and research on the structure of human abilities. Psychological Inquiry, 3, 365-384.

Nicholson, C. L. (1992, Fall). CDES Communique (Available from Council for Educational Diagnostic Services, Council for Exceptional Children, 1920 Association Drive, Reston, Virginia 22091).

Osterlind, S. J. (1989). Constructing test items. Boston: Kluwer Academic Publishers.

Page, W.D. (1977). Comprehension and Cloze Performance. Reading World, 17, 1-12.

Palincsar, A.S., & Perry, N.E. (1995). Developmental, cognitive, and sociocultural perspectives on assessing and instructing reading. School Psychology Review, 24(3), 331-344.

Perfetti, C.A.(1994). Reading. In R.J. Sternberg (Ed.), Encyclopedia of human intelligence (pp. 923-930). New York: McMillian Publishing Company.

Prewett, P. N. & Giannuli, M. M. (1991). The relationship among the reading subtests of the WJ-R, PIAT-R, K-TEA, and WRAT-R. Journal of Psychoeducational Assessment, 9, 166-174.

Rankin, E.F. (1958). An evaluation of the cloze procedure as a technique for measuring reading comprehension. Dissertation Abstracts International, 19, 733-734. University Microfilms No. 58-3722.

Rankin, E. F. & Culhane, J. (1969). Comparable cloze and multiple choice comprehension test scores. Journal of Reading 13, 193-198.

Salvia, J., & Ysseldyke, J.E. (1991). Assessment in special and remedial education (5th ed.). Boston: Houghton-Mifflin.

Slate, J. R. (1996). Interrelations of frequently administered achievement measures in the determination of specific learning disabilities. Learning Disabilities Research and Practice, 11, 86-89.

Smith, N. & Zinc, A. (1977). A cloze-based investigation of reading comprehension as a composite of subskills. Journal of Reading Behavior 9, 395-398.

Spearitt, D.(1972). Identification of subskills of reading comprehension by maximum likelihood factor analysis. Research Quarterly, 8, 92-111.

Stahl, S. A. & Murray, B. A. (1994). Defining phological awareness and its relatationship to early reading. Journal of Educational Psychology, 86, 221-234.

Stinnett, T.A., Havey, J.M., & Oehler-Stinnett, J. (1994). Current test usage by practicing school psychologists: A national survey. Journal of Psychoeducational Assessment, 12, 331-350.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. Journalism Quarterly, 30, 415-433.

Thompson, S. S. (1993). Test review: Wechsler Individual Achievement Test (WIAT). Journal of Psychoeducational Assessment, 11, 292-297.

Thorndike, R.L.(1974). Reading as reasoning. Reading Research Quarterly, 9, 137-147.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1994). Longitudinal studies of phonological processing and reading. Journal of Learning Disabilities, 27(5), 276-286

Wagner, R.K., & Torgesen, J.K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. Psychological Bulletin, 101(2), 192-212.

Wagner, R.K., Torgesen, J.K., Laughon, P., Simmons, K., & Rashotte, C.A. (1993). Development of young readers' phonological processing abilities. Journal of Educational Psychology, 85(1), 83-103.

Ward, A.W. (1995). Review of the Wide Range Achievement Test 3. In J.C. Conoley, & J.C. Impara (Eds.), The twelfth mental measurements yearbook (pp. 1110-1111). Lincoln, NE: Buros Institute.

Wiederholt, J.L. (1986). Formal Reading Inventory. Austin, TX: PRO-ED.

Wiederholt, J.L. & Bryant, B.R. (1986). Gray Oral Reading Tests-Revised (GORT-R). San Antonio, TX: The Psychological Corporation.

Wilkinson, G. S. (1993). Wide Range Achievement Test 3. Wilmington, DE: Wide Range, Inc.

Woodcock, R. W. (1987). Woodcock Reading Mastery Test--Revised. Circle Pines, MN: American Guidance Service.

Woodcock, R.W. (1990). Theoretical foundations of the WJ-R measures of cognitive ability. Journal of Psychoeducational Assessment, 8, 231-258.

Woodcock, R. W. (1997). Woodcock Diagnostic Reading Battery. Chicago: Riverside.

Woodcock, R. W. (in press). Extending Gf-Gc theory into practice. In McArdle, J. J., & Woodcock, R. W. (Eds.), Human cognitive abilities in theory and practice. Chicago: Riverside.

Woodcock, R.W. & Johnson, M.B. (1989). Woodcock-Johnson Psychoeducational Battery-Revised. Chicago: Riverside.

Woodcock, R.W., McGrew, K.S., & Werder, J.K. (1994). Woodcock-McGrew-Werder Mini-Battery of Achievement. Chicago: Riverside.

Appendix A

Path Diagrams for Final *Gf-Gc* Cross-Battery Confirmatory Factor Analyses for Four Samples

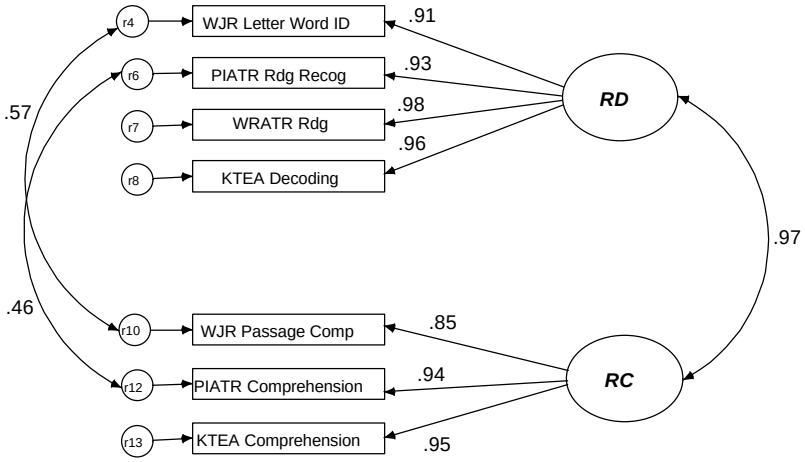
Prewett & Giannuli (1991) Sample

WJ-R Grade 3 Validity Sample

WJ-R Grade 3/4 Validity Sample

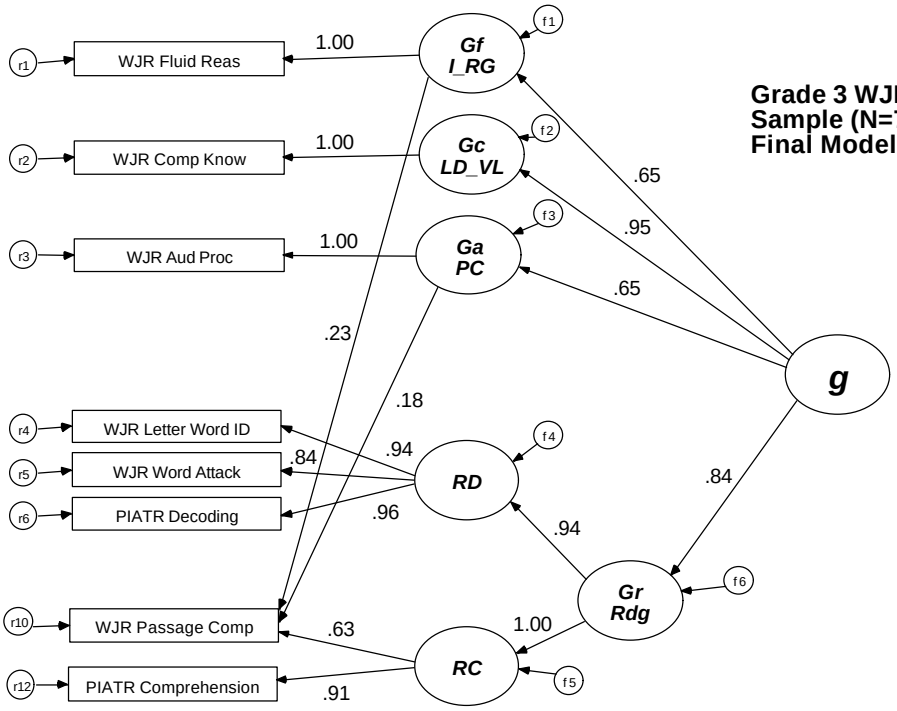
WJ-R Grade 10/11 Validity Sample

Prewett & Giannuli (1991)
Sample (n=118)
Final Model



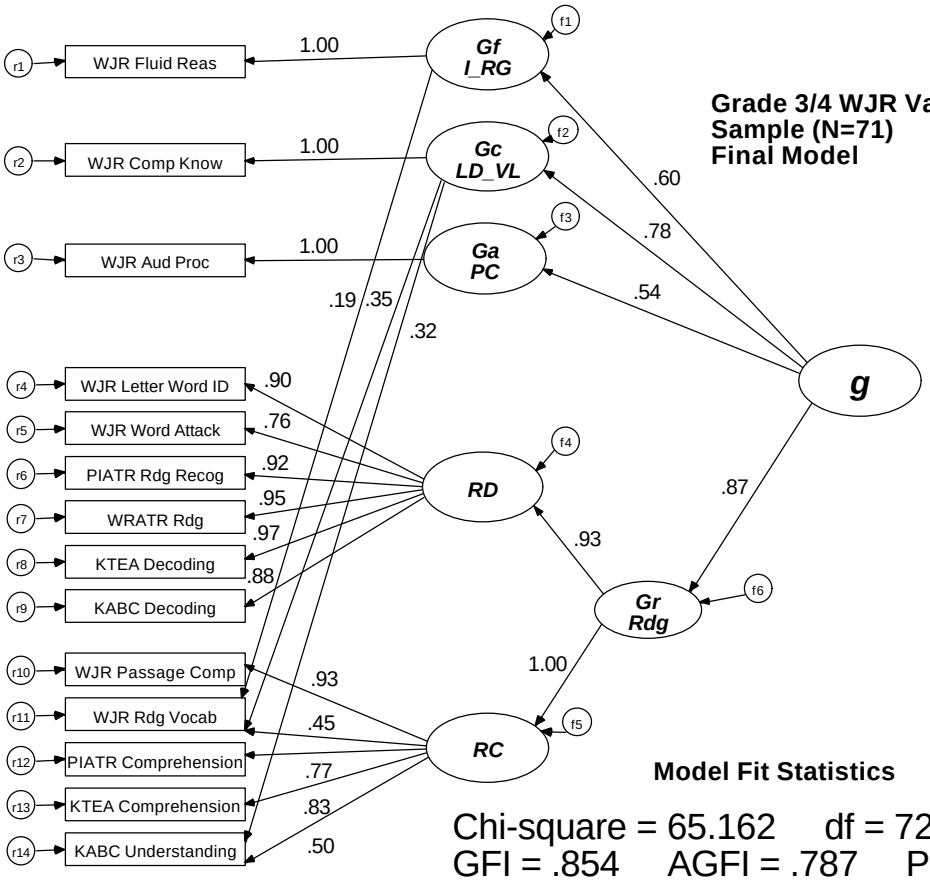
Model Fit Statistics

Chi-square = 54.662 df = 11 p = .000
GFI = .893 AGFI = .727 PGFI = .351
CFI = .967 PCFI = .507 rmsea = .184



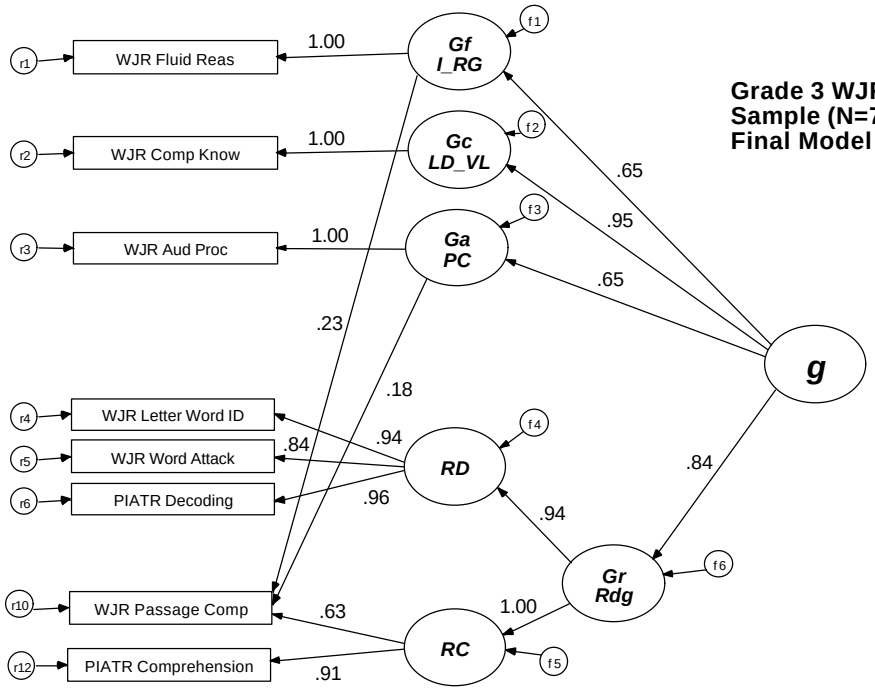
Model Fit Statistics

Chi-square = 47.045 df = 16 p = .000
GFI = .862 AGFI = .690 PGFI = .383
CFI = .943 PCFI = .539 rmsea = .164



Model Fit Statistics

Chi-square = 65.162 df = 72 p = .703
GFI = .854 AGFI = .787 PGFI = .586
CFI = 1.000 PCFI = .791 rmsea = .000

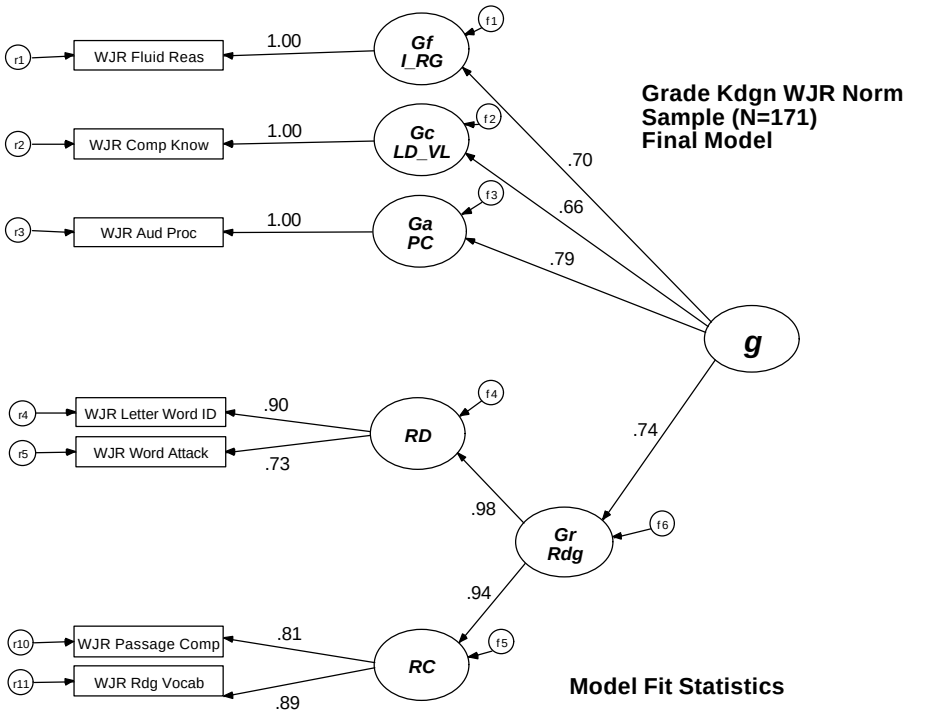


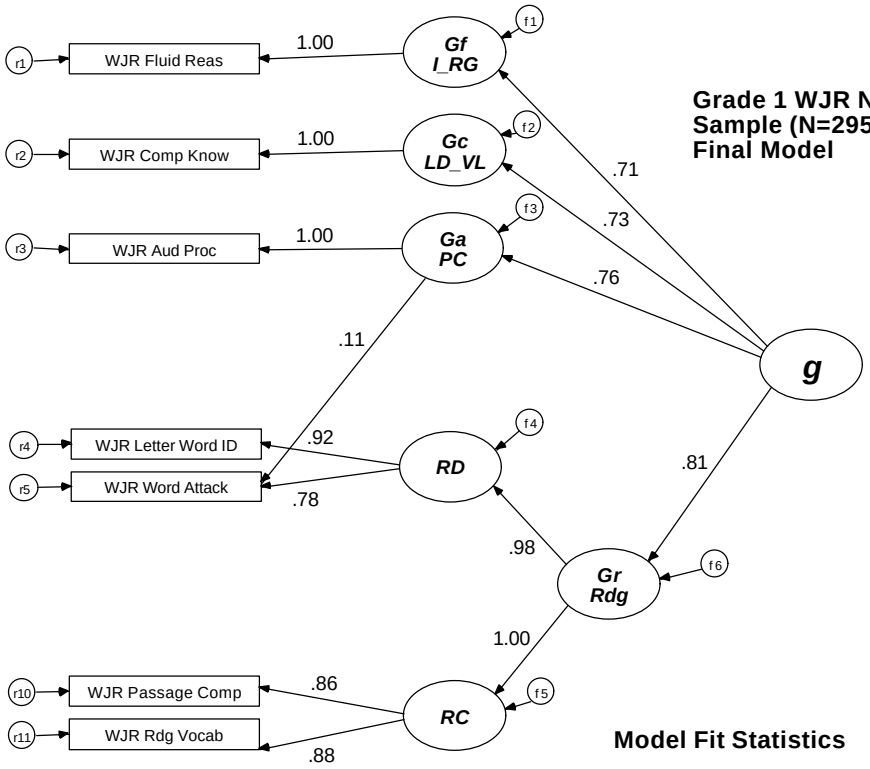
Model Fit Statistics

Chi-square = 47.045 df = 16 p = .000
GFI = .862 AGFI = .690 PGFI = .383
CFI = .943 PCFI = .539 rmsea = .164

Appendix B

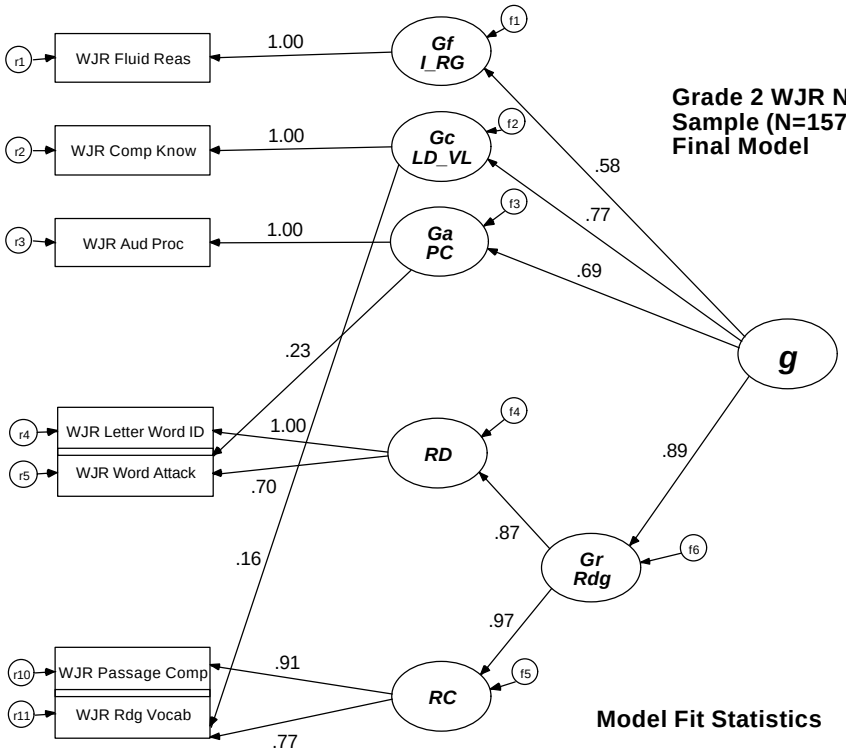
Path Diagrams for Final *Gf-Gc* Confirmatory Factor Analyses in Four WJ-R Norm Samples (Kdgn - 3rd Grade)


Grade Kdgn WJR Norm
Sample (N=171)
Final Model
r1
WJR Fluid Reas
1.00
Gf
I_RG
f1
r2
WJR Comp Know
1.00
Gc
LD_VL
f2
r3
WJR Aud Proc
1.00
Ga
PC
f3
.70
.66
.79
g
r4
WJR Letter Word ID
r5
WJR Word Attack
.90
.73
RD
f4
.74
.98
Gr
Rdg
f6
.94
r10
WJR Passage Comp
r11
WJR Rdg Vocab
.81
.89
RC
f5
Model Fit Statistics
Chi-square = 18.986 df = 11 p = .061
GFI = .968 AGFI = .919 PGFI = .380
CFI = .986 PCFI = .517 rmsea = .065



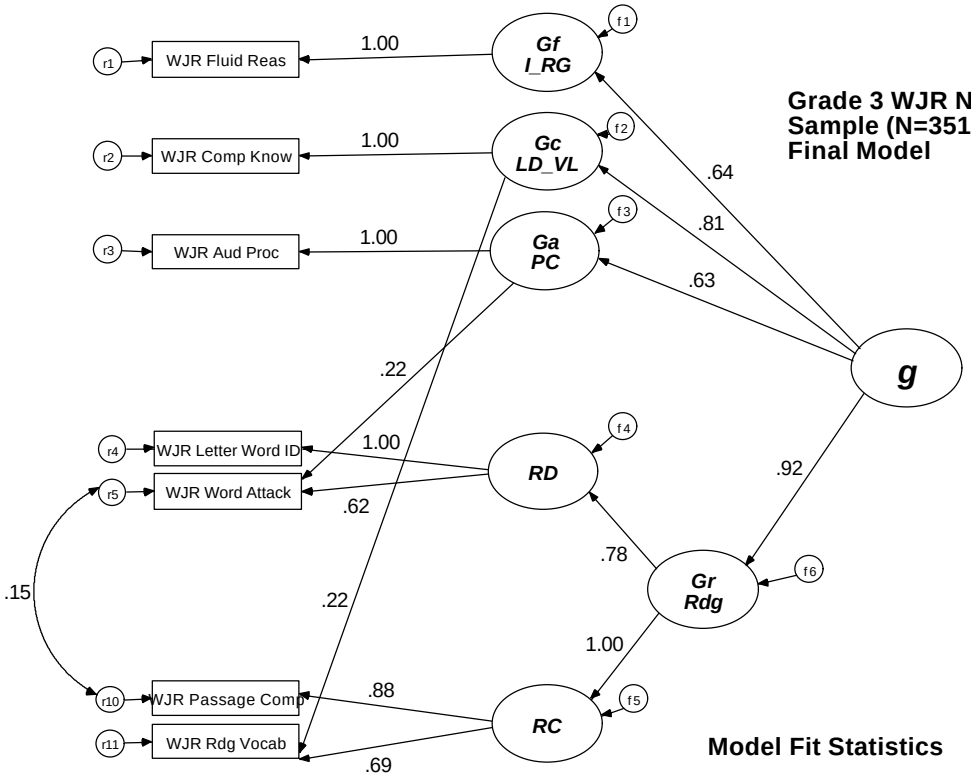
Model Fit Statistics

Chi-square = 50.451 df = 11 p = .000
GFI = .953 AGFI = .882 PGFI = .375
CFI = .971 PCFI = .509 rmsea = .110



Model Fit Statistics

Chi-square = 14.458 df = 10 p = .153
GFI = .974 AGFI = .927 PGFI = .348
CFI = .994 PCFI = .473 rmsea = .053


Grade 3 WJR Norm Sample (N=351) Final Model
WJR Fluid Reas
WJR Comp Know
WJR Aud Proc
WJR Letter Word ID
WJR Word Attack
WJR Passage Comp
WJR Rdg Vocab
Gf I_RG
Gc LD_VL
Ga PC
RD
Gr Rdg
RC
g
1.00
.64
.81
.63
.22
.92
.62
.78
.15
.88
.69
Model Fit Statistics
Chi-square = 13.776 df = 9 p = .131
GFI = .989 AGFI = .966 PGFI = .318
CFI = .997 PCFI = .427 rmsea = .039