

Are Cattell–Horn–Carroll Broad Ability Composite Scores Exchangeable Across Batteries?

Randy G. Floyd
Renee Bergeron
Allison C. McCormack
Janice L. Anderson
Gabrielle L. Hargrove-Owens
The University of Memphis

Abstract. Many school psychologists use the Cattell–Horn–Carroll (CHC) theory of cognitive abilities to guide their interpretation of scores from intelligence test batteries. Some may frequently assume that composite scores purported to measure the same CHC broad abilities should be relatively similar for individuals no matter what subtests or batteries were administered to obtain these scores. This study examined this assumption using six samples of preschool children, school-age children, or adults who completed two or more intelligence test batteries. From these samples, composites measuring the broad abilities Crystallized Intelligence, Visual Processing, Fluid Reasoning, and Processing Speed were compared to examine their exchangeability. Results indicate that most CHC broad ability composites produced scores that were not as exchangeable for individuals as may have been assumed by some. Discussion focuses on the influence of score reliability and on the interaction between examinee characteristics and the tasks used to measure the broad abilities.

Two notable trends in the assessment of cognitive abilities have emerged. The first is the movement toward theory-based test development and test interpretation (Kamphaus, Winsor, Rowe, & Kim, 2005). Perhaps most prominent among these theories is the Cattell–Horn–Carroll (CHC) theory of cog-

nitive abilities (Alfonso, Flanagan, & Radwan, 2005). Many recognize the sources of this theory, the Cattell–Horn *G_f–G_c* theory (Horn, 1991; Horn & Blankson, 2005) and the Carroll three-stratum theory (Carroll, 1993, 2003), as the most complete and empirically supported descriptions of the structure of

We thank the Woodcock–Munoz Foundation, Richard Woodcock, Fredrick Schrank, and Kevin McGrew as well as American Guidance Service, Marshall Dahl, and Scott Overgaard for providing data for this study. We also thank those who we know coordinated data collection: Laurie Ford, Terri Teague, MB Tusing, Susan League, LeAdelle Phelps, David McIntosh, Mardis Dunham, Noel Gregg, and Cheri Hoy. We thank Jaime Hart for data organization and Kevin McGrew for comments about this study. Tom Fagan, Sam Ortiz, and Dawn Flanagan provided useful feedback about earlier drafts. We are also appreciative of information from Bruce Bracken and Larry Evans. Portions of this research were presented at the annual meeting of the American Psychological Association (2004) and the National Association of School Psychologists (2005).

Address correspondence to Randy G. Floyd, The University of Memphis, Department of Psychology, Memphis, TN 38152; E-mail: rgfloyd@memphis.edu

Copyright 2005 by the National Association of School Psychologists, ISSN 0279-6015

human cognitive abilities (Kamphaus, 2001; McGrew, 2005; Sattler, 2001).

CHC theory describes a hierarchical model of cognitive abilities that vary according to level of generality: narrow abilities (Stratum I), broad abilities (Stratum II), and in the minds of some, general intelligence (g ; Stratum III). Narrow abilities include approximately 70 highly specialized abilities. Broad abilities include Fluid Reasoning, Crystallized Intelligence, Short-Term Memory, Visual Processing, Auditory Processing, Long-Term Retrieval, Processing Speed, Reading and Writing Ability, Quantitative Knowledge, and Reaction Time/Decision Speed. There appears to be consensus that the CHC theory provides a common nomenclature for describing the broad abilities measured across test batteries (see Flanagan & Ortiz, 2001; McGrew & Flanagan, 1998). In contrast, the existence of a single higher order general factor (g) is the focus of much debate—even among namesakes of the CHC theory (see McGrew, 2005). Some researchers, such as Carroll (1993, 2003) and Jensen (1998), assert that (a) this general factor represents well what is shared among the broad abilities and (b) it is the only cognitive ability tapped by all ability measures. Conversely, others researchers, such as Horn (1991), argue for a focus on the somewhat independent broad abilities and relegate g to a Protean and relatively meaningless conglomerate of more specific cognitive abilities. From this latter perspective, measures of most or all of CHC broad abilities should be considered, with relatively equal weight, when completing a cognitive ability assessment (Carroll, 1993).

The influence of what would become CHC theory gained momentum in school psychology in the late 1980s and early 1990s with the publication of the Woodcock–Johnson Tests of Cognitive Ability, Revised (Woodcock & Johnson, 1989) and the publication of Woodcock’s (1990) research examining broad abilities measured by prominent intelligence tests. Later in the 1990s, McGrew, Flanagan, and Ortiz published works espousing the Cross-Battery approach to guide the comprehensive measurement of broad and narrow abilities (Flanagan & McGrew, 1997;

Flanagan, McGrew, & Ortiz, 2000; Flanagan & Ortiz, 2001; McGrew & Flanagan, 1998). Subsequently, test authors and publishers used CHC theory to guide the revisions of several prominent test batteries (see Alfonso et al., 2005). These batteries include the Woodcock–Johnson III Tests of Cognitive Abilities (WJ III; Woodcock, McGrew, & Mather, 2001), the Stanford–Binet Intelligence Scales, Fifth Edition (SB5; Roid, 2003), and the Kaufman Assessment Battery for Children, Second Edition (KABC-II; Kaufman & Kaufman, 2004a). The Wechsler Intelligence Scale for Children, Fourth Edition (Wechsler, 2003) also appears to be more closely aligned with CHC theory than its predecessors. Based on this apparent influence of CHC theory, Flanagan and Kaufman (2004) remarked, “Never before in the history of intelligence testing has a single theory (indeed any theory) played so prominent a role in test development and interpretation” (p. 14).

The second trend in the assessment of cognitive abilities has stemmed directly from CHC theory and its applications to test interpretation. Based on the salient criticisms of interpreting *subtests* as measures of specific cognitive abilities (e.g., McDermott, Fantuzzo, & Glutting, 1990), interpretation of measures of cognitive abilities appears to have focused on composite scores designed to represent broad and narrow abilities with greater reliability than subtests. The Cross-Battery approach of McGrew, Flanagan, and colleagues provided the means to operationalize nine of the CHC broad abilities and numerous narrow abilities via composite scores. Consistent with this trend, the WJ III, SB5, and KABC-II all produce composites measuring Fluid Reasoning, Crystallized Intelligence, Short-Term Memory, and Visual Processing. The WJ III and the KABC-II produce composites measuring Long-Term Retrieval. Research using some of these CHC broad ability composites has focused on their relative importance in predicting achievement domains (e.g., Evans, Floyd, McGrew, & Leforgee, 2002; Floyd, Evans, & McGrew, 2003) and on profiles of these composites for children with exceptionalities (e.g., Floyd, Bergeron, & Alfonso, 2005; Proctor,

Floyd, & Shaver, 2005; Rizza, McIntosh, & McCunn, 2001). Based on these two trends, it is apparent that interpretation of composite scores representing the CHC broad abilities is the focus of notable research attention, and it very likely has become the focus of cognitive ability assessment by many school psychologists.

Score Exchangeability

There are several issues that deserve considerable contemplation and research when considering these two trends (Glutting, Watkins, & Youngstrom, 2003). One issue relates to the assumption that different composite scores—purported to measure the same ability—from different batteries provide comparable norm-based scores for individuals. Just as consumers expect that thermometers sold by different manufacturers, blood pressure gauges developed by different medical suppliers, and radar guns that came from different vendors should yield similar temperatures, provide similar readings of systolic and diastolic blood pressure, or similar vehicle speeds on the same target, many assume that measures of the same construct should, within reason, yield similar results on the same individual. This issue is best described as one of *score exchangeability* (Floyd, Clark, & Shadish, 2005). Score exchangeability is consistent with the well-known psychometric property *convergent validity* (Campbell & Fiske, 1959), which focuses on patterns of relations between measures of the same construct measured by different methods, and similar to the goal of *convergence of indicators* (Messick, 1989), which focuses on the assumption that individuals should perform at similar levels on a variety of indicators of a construct. Methods to examine score exchangeability focus on the level of analysis of the individual in a manner mirroring the day-to-day decisions of school psychologists and others involved in ability assessments who strive to generalize their interpretation of these composites to families of related measures available to them. This focus on the individual is consistent with calls for a *person-oriented* approach and accompanying methodology to understand individual patterns of development over time (Bergman

& Magnusson, 1997; Cairns, Bergman, & Kagan, 1998). It also is in response to apparent frustration with the poor ability of group-data methodology to focus on the response patterns of individuals on intelligence tests (e.g., Flanagan & Kaufman, 2004).

Score exchangeability is probably best examined using a number of statistical techniques. In research and test manuals, it is typical that correlations between two scores purported to measure the same construct are presented. These results are often accompanied by reports of the difference between the average scores (i.e., means) from the two measures included in the correlational analyses.¹ For example, in a sample of 8- to 12-year-old children, the correlation between the KABC Mental Processing Composite (Kaufman & Kaufman, 1983) and the KABC-II Fluid-Crystallized Index was shown to be .81, and the difference in the means for two scores was 5.3 standard score points (Kaufman & Kaufman, 2004a, p. 110). However, correlations and difference in mean values convey only how scores tend to agree when ranking individuals above and below the average score on respective measures and whether the average score for one measure tends to be higher than the average score for another measure. Individually, each of these statistical techniques directly conveys little about what should be expected at the level of the individual.

Additional statistical techniques can focus greater attention both on unreliability of measurement and on absolute differences between scores at the level of the individual. In order to consider unreliability of measurement in examining score exchangeability, confidence intervals (based on reliability coefficients) surrounding obtained scores can be considered (Charter & Feldt, 2000). With this consideration, the meaningfulness of differences between obtained scores can be judged more appropriately. Because typical correlation coefficients neither quantify the extent to which two scores differ on an absolute level in measuring the same ability nor allow for more than two scores to be compared, *absolute unit dependability coefficients* using generalizability theory can be

reported (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). For example, the absolute unit dependability coefficients measure the extent to which a single score can be generalized to the universe of scores measuring the same construct. In addition, generalizability theory analyses can provide information about the influences that have the largest effects on scores. Overall, these techniques examining exchangeability bring the researcher closer to the level of the individual during the consideration of score differences.

Purpose of the Study

This study was designed to (a) examine the assumption of score exchangeability when considering CHC broad ability composites and (b) yield results reflecting the reasons for differences in exchangeability values. To investigate these issues, 40 composite scores measuring four broad abilities most commonly included in prominent test batteries—Fluid Reasoning, Crystallized Intelligence, Visual Processing, and Processing Speed—were compared to other composites measuring the same ability. Composites were formed from two or more subtests² from seven contemporary intelligence test batteries. Exchangeability statistics were used to compare how well the composite scores produce similar scores for individuals.

Method

Participants

Composite scores stemmed from data from six samples of children or adults that were collected as part of the process designed to produce validity evidence supporting the use of published cognitive ability test batteries. Across all samples, more than 800 children and adults participated. They ranged from 3 years to more than 50 years of age.

Sample 1. As reported by McGrew and Woodcock (2001), 202 toddler- and preschool-age children completed in counterbalanced order the Wechsler Preschool and Primary Scale of Intelligence—Revised (WPPSI-R; Wechsler, 1989), the Differential Ability Scales

(DAS; Elliott, 1990), and the WJ III (Woodcock et al., 2001). From this sample, children ages 36 months and older were selected. The subsample included 117 White children (65.4%), 59 Black children (33.0%), and 3 Asian/Pacific Islander children (1.7%).

Sample 2. As reported by McGrew and Woodcock (2001) and Phelps, McGrew, Knopik, and Ford (2005), 150 children in Grades 3, 4, and 5 completed in counterbalanced order the Wechsler Intelligence Scale for Children—Third Edition (WISC-III; Wechsler, 1991) and the WJ III (Woodcock et al., 2001). The sample included 148 White children (98.7%).

Sample 3. As reported by McGrew and Woodcock (2001), 130 children in Grades 3, 4, and 5 completed in counterbalanced order the DAS (Elliott, 1990) and the WJ III (Woodcock et al., 2001). The sample included 124 White children (95.4%) and 6 Black children (4.6%).

Sample 4. As reported by Kaufman and Kaufman (2004a), 119 children ages 8 to 13 ($M = 124.7$ months, $SD = 18.5$ months) completed in counterbalanced order the KABC-II (Kaufman & Kaufman, 2004a) and the WISC-III (Wechsler, 1991). From this sample, children ages 8 to 12 were selected because the battery differs somewhat for children ages 13 and older. The sample included 71 White children (61.2%), 18 Hispanic children (15.5%), 9 African American children (7.8%), and 9 Asian children (7.8%).

Sample 5. As reported by Kaufman and Kaufman (2004a), 86 children ages 7 to 16 completed in counterbalanced order the KABC-II (Kaufman & Kaufman, 2004a) and the WJ III (Woodcock et al., 2001). From this sample, children ages 8 to 12 were selected. The sample included 50 White children (60.2%), 16 Hispanic children (19.3%), 8 Asian children (9.3%), and 5 African American children (6%).

Sample 6. As reported by McGrew and Woodcock (2001), 205 undergraduate students completed in counterbalanced order (a) the

WAIS-III (Wechsler, 1997) and the WJ III (Woodcock et al., 2001), (b) the Kaufman Adolescent and Adult Intelligence Test (KAIT; Kaufman & Kaufman, 1993) and the WJ III, or (c) all three test batteries. About half of the sample were students diagnosed with a learning disability or a learning disability and attention-deficit/hyperactivity disorder. The other half volunteered as participants. The sample included 187 White students (92.1%) and 14 Black students (6.9%).

Measures

Composites are organized in Table 1 by the CHC broad abilities they measure. The table presents descriptions and psychometric properties of 40 composites used in the analyses. Subtests contributing to each composite are listed, and these subtests are bolded if factor analytic evidence organized in a manner consistent with CHC theory supports the broad ability designation. The primary CHC narrow ability thought to be measured by each subtest is also listed and marked with an asterisk if more than one narrow ability may be measured (Flanagan & Ortiz, 2001; Kaufman & Kaufman, 2004a; McGrew & Flanagan, 1998; McGrew & Woodcock, 2001).

As evident in Table 1, 13 composites stemmed directly from scoring the test battery (*test battery-based composites*), and 27 were derived from two or more norm-based subtest scores (*derived composites*). Test battery-based composites were included if they contained at least two subtests that yielded qualitatively different measures of CHC narrow abilities.³ Derived composites were constructed when test batteries (a) contained at least two subtests measuring the same broad ability and different narrow abilities⁴ and (b) did not include them in a test battery-based composite consistent with CHC theory.⁵ When more than two subtests measuring the same broad ability were identified within a test battery, all pair-wise combinations of those subtests were used to create composites. To form derived composites, scaled scores ($M = 10$, $SD = 3$) from the WPPSI-R, WISC-III, KABC-II, WAIS-III, and KAIT, and T scores ($M = 50$, $SD = 10$) from the DAS were first converted to devia-

tion IQ scores ($M = 100$, $SD = 15$). Relevant subtest scores were then averaged to produce broad ability composites. All scores stemmed from use of age-based norms.

Reliability. Reliability coefficients for test battery-based and derived composites were obtained from information included in test battery manuals, and they are reported in Table 1. All reported reliability coefficients, except those for the three Processing Speed composites, are internal consistency coefficients, such as split-half reliability coefficients. The reliability coefficients for the Processing Speed composites are test-retest reliability coefficients. Reliability coefficients for test battery-based composites are (a) either means or medians reported in their respective test manuals or (b) means across specified age groups calculated for this study. Reliability coefficients for derived composites were calculated using (a) subtest reliability coefficients averaged across ages and (b) correlations between subtests (Nunnally & Bernstein, 1994). When the test manual did not report a mean or median reliability coefficient across age groups with the level of specificity needed for the analysis, the average (mean) reliability for test-based composites or for subtests in a derived composite was calculated by converting the reliability coefficients for each specific age group to z scores using Fisher's transformation, averaging the z scores, and converting the average z score back to a coefficient. In every case in which correlations between subtests were needed to calculate reliability estimates for derived composites, correlations for the specified age groups were reported in the manuals (i.e., WPPSI-R, WISC-III, DAS, WJ III, and KABC-II).

General factor loadings. General factor loadings represent the correlations between the subtests and the latent factor that is presumed to be g . These values were first obtained for each subtest contributing to a composite, and they are reported in Table 1. Two methods were employed. First, g loadings were obtained from test manuals, test authors, or other published resources (Elliott, 1990; Kaufman & Kaufman, 1993, 2004a; K. S. McGrew, per-

Table 1
Descriptions and Psychometric Properties of Broad Ability Composites

Composite Label	Type	Subtests in Composite ^a	CHC Narrow Ability for Subtests	Reliability ^b	Subtest <i>g</i> Loadings ^c	Composite <i>g</i> Loading
<i>Comprehension-Knowledge (Gc)</i>						
WPPSI-R Gc4	D	Vocabulary	Lexical Knowledge		.71/.60	
		Similarities	Language Development ^d	.94	.72/.68	.74/.70
		Comprehension	Language Development ^d		.72/.73	
		Information	General Information		.79/.78	
WPPSI-R Gc2:1	D	Vocabulary	Lexical Knowledge	.90	.71/.60	.72/.64
		Similarities	Language Development ^d		.72/.68	
WPPSI-R Gc2:2	D	Vocabulary	Lexical Knowledge	.90	.71/.60	.72/.67
		Comprehension	Language Development ^d		.72/.73	
WPPSI-R Gc2:3	D	Vocabulary	Lexical Knowledge	.90	.71/.60	.75/.69
		Information	General Information		.79/.78	
WPPSI-R Gc2:4 ^e	D	Similarities	Language Development ^d	.90	.72/.68	.72/.71
		Comprehension	Language Development ^d		.72/.73	
WPPSI-R Gc2:5	D	Similarities	Language Development ^d	.93	.72/.68	.76/.73
		Information	General Information		.79/.78	

(Table 1 continues)

(Table 1 continued)

Composite Label	Type	Subtests in Composite ^a	CHC Narrow Ability for Subtests	Reliability ^b	Subtest g Loading ^c	Composite g Loading
WPPSI-R Gc2:6	D	Comprehension Information	Language Development ^d General Information	.90	.72/.73 .79/.78	.76/.76
DAS Verbal	T	Verbal Comprehension Naming Vocabulary	Language Development ^d Lexical Knowledge ^d	.88	.72/.78 .64/.76	.68/.77
WJ III	T	Verbal	Lexical Knowledge ^d		.81/.69 .79/.77	
Comprehension–Knowledge (Gc)		Comprehension General Information	General Information	.93 .94 .70/.68 .74/.72		.76/.69 .77/.75
WISC–III Verbal	T	Vocabulary	Lexical Knowledge		.79/.77	
Comprehension Index (VCI)		Similarities Comprehension	Language Development ^d Language Development ^d	.94	.78/.80 .70/.61	.76/.74
WISC–III Gc2:1	C	Information Vocabulary	General Information Lexical Knowledge	.91	.78/.76 .79/.77	.79/.79
WISC–III Gc2:2	C	Similarities Vocabulary	Language Development ^d Lexical Knowledge	.89	.78/.80 .79/.77	.75/.69
		Comprehension	Language Development ^d		.70/.61	

(Table 1 continues)

(Table 1 continued)

Composite Label	Type	Subtests in Composite ^a	CHC Narrow Ability for Subtests	Reliability ^b	Subtest g Loading ^c	Composite g Loading
WISC-III Gc2:3	C	Vocabulary	Lexical Knowledge	.91	.79/.77	.79/.77
		Information	General Information		.78/.76	
WISC-III Gc2:4 ^e	C	Similarities	Language Development ^d	.87	.78/.80	.74/.71
		Comprehension	Language Development ^d		.70/.61	
WISC-III Gc2:5	C	Similarities	Language Development ^d	.89	.78/.80	.78/.78
		Information	General Information		.78/.76	
WISC-III Gc2:6	C	Comprehension	Language Development ^d	.88	.70/.61	.74/.69
		Information	General Information		.78/.76	
KABC-II Gc	T	Riddles	Lexical Knowledge ^d	.92	.78/.72	.77/.74
		Verbal Knowledge	Lexical Knowledge ^d		.76/.76	
KABC-II Gc2:1	D	Riddles	Lexical Knowledge ^d		.78/.72	
		Expressive	Lexical Knowledge	.92	.71/.74	.75/.73
		Vocabulary				
KABC-II Gc2:2 ^e	D	Verbal Knowledge	Lexical Knowledge ^d		.76/.76	
		Expressive	Lexical Knowledge	.92	.71/.74	.74/.75
		Vocabulary				

(Table 1 continues)

(Table 1 continued)

Composite Label	Type	Subtests in Composite ^a	CHC Narrow Ability for Subtests	Reliability ^b	Subtest <i>g</i> Loadings ^c	Composite <i>g</i> Loading
<i>Visual Processing (Gv)</i>						
WPPSI-R Gv3	D	Block Design	Spatial Relations ^d		.68/.69	
		Object Assembly	Closure Speed ^d	.87	.57/.62	.60/.62
		Mazes	Spatial Scanning		.56/.54	
WPPSI-R Gv2:1	D	Block Design	Spatial Relations ^d	.83	.68/.69	.63/.66
		Object Assembly	Closure Speed ^d		.57/.62	
WPPSI-R Gv2:2	D	Block Design	Spatial Relations ^d	.87	.68/.69	.62/.62
		Mazes	Spatial Scanning		.56/.54	
WPPSI-R Gv2:3	D	Object Assembly	Closure Speed ^d	.78	.57/.62	.57/.58
		Mazes	Spatial Scanning		.56/.54	
DAS Gv2	D	Pattern Construction	Spatial Relations	.88	.59/.69	.54/.60
		Recognition of Pictures	Visual Memory		.49/.50	
WJ III Visual–	T	Spatial Relations	Visualization, Spatial		.65/.63	.47/.31
Spatial Thinking			Relat.	.88	.79	.60/.56
(Gv)		Picture Recognition	Visual Memory		.55/.48	.39/.38

(Table 1 continues)

(Table 1 continued)

Composite Label	Type	Subtests in Composite ^a	CHC Narrow Ability for Subtests	Reliability ^b	Subtest <i>g</i> Loading ^c	Composite <i>g</i> Loading
WISC-III Gv2	D	Block Design	Spatial Relations ^d	.86	.74/.59	.70/.55
		Object Assembly	Closure Speed ^d		.66/.51	
DAS Spatial	T	Recall of Designs	Visual Memory	.92	.63/.57	.67/.57
		Pattern Construction	Spatial Relations		.70/.56	
KABC-II Gv	T	Rover	Spatial Scanning ^d	.88	.54/.55	.59/.60
		Triangles	Spatial Relations ^d		.63/.65	
KABC-II Gv2:1	D	Rover	Spatial Scanning ^d	.87	.54/.55	.54/.51
		Block Counting	Visualization ^d		.54/.46	
KABC-II Gv2:2	D	Rover	Spatial Scanning ^d	.80	.54/.55	.52/.42
		Gestalt Closure	Closure Speed		.49/.28	
KABC-II Gv2:3	D	Triangles	Spatial Relations ^d	.90	.63/.65	.59/.56
		Block Counting	Visualization ^c		.54/.46	
KABC-II Gv2:4	D	Triangles	Spatial Relations ^d	.85	.63/.65	.51/.47
		Gestalt Closure	Closure Speed		.49/.28	
KABC-II Gv2:5	D	Block Counting	Visualization ^d	.83	.54/.46	.52/.37
		Gestalt Closure	Closure Speed		.49/.28	

(Table 1 continues)

(Table 1 continued)

Composite Label	Type	Subtests in Composite ^a	CHC Narrow Ability for Subtests	Reliability ^b	Subtest g Loading ^c	Composite g Loading
<i>Fluid Reasoning (Gf)</i>						
DAS Nonverbal	T	Matrices	Induction		.71/.70	
		Sequential and Quantitative Reasoning	Quantitative Reasoning ^d	.90	.76/.71	.74/.71
		Concept Formation	Induction		.73/.66	.76/.69
WJ III Fluid Reasoning	T	Analysis–Synthesis	General Sequential	.95	.67/.77	.71/.61
			Reason.			.70/.69
		Pattern Reasoning	Induction ^e	.89	.66/.66	.63/.62
		Story Completion	Induction ^e		.60/.58	
		Logical Steps	General Sequential		.72/.66	
		Mystery Codes	Reason.	.93		.71/.68
			Induction		.69/.70	

(Table 1 continues)

(Table 1 continued)

Composite Label	Type	Subtests in Composite ^a	CHC Narrow Ability for Subtests	Reliability ^b	Subtest <i>g</i> Loading ^c	Composite <i>g</i> Loading
<i>Processing Speed (Gs)</i>						
WISC-III	T	Coding	Rate-of-Test-Taking		.44/.42	
Processing Speed Index		Symbol Search	Perceptual Speed ^d	.85	.62/.54	.53/.48
				E	A	E
				A	A	A
WJ III Processing Speed (Gs)	T	Visual Matching Decision Speed	Perceptual Speed Semantic Processing	.91	.53/.52 .49/.41	.56/.36 .51/.48
						.51/.47
						.54/.42
			Speed			
WAIS-III	T	Digit Symbol-Coding	Rate-of-Test-Taking		.59/.31	
Processing Speed Index		Symbol Search	Perceptual Speed ^d	.87	.70/.38	.65/.35

Note: D = derived composite; T = test battery-based composite; P = preschool age; E = elementary school age; A = adult age; DAS = Differential Abilities Scales; KAIT = Kaufman Adolescent and Adult Intelligence Test; KABC-II = Kaufman Assessment Battery for Children—Second Edition; WAIS-III = Wechsler Adult Intelligence Scale—Third Edition; WISC-III = Wechsler Intelligence Scale for Children—Third Edition; WPPSI-R = Wechsler Preschool and Primary Scale of Intelligence, Revised; WJ III = Woodcock-Johnson III Tests of Cognitive Abilities.

^aSubtests are bolded if factor analytic evidence organized in a manner consistent with CHC theory supports the broad ability classification. ^bReliability coefficients for test battery-based composites and for subtests used in derived composites were averages across all ages as reported in their respective manuals; WPPSI-R (ages 3 to 7), WISC-III (6 to 16), and WAIS-III (16 to 89), and KAIT (11 to 89+). The reliability coefficients for the WISC-III Processing Speed Index and WAIS-III Processing Speed Index were medians across their respective age ranges. For the DAS, reliability coefficients for test-based composites and for subtests used in derived composites were averages across (a) ages 3;6 to 5;11 for the preschool battery (for Sample 1)

(Table 1 continues)

(Table 1 continued)

or (b) ages 6 to 17:11 for the school-age battery (for Sample 3). For the WJ III, reliability coefficients for test-based composites were averages across (a) ages 4 to 5 for the comparisons at the preschool level (Sample 1), (b) ages 9 to 13 for the elementary school-age comparisons (for Samples 2, 3, and 5), (c) and ages 20 to 39 for the adult comparisons (for Sample 6). For the KABC-II, reliability coefficients for test-based composites and for subtests used in derived composites were averages across ages 7 to 12. Sattler (2001), Sattler and Ryan (2001) and Sattler and Saklofske (2001) reported *g* loadings using principal components analysis (PCA) for the WPPSI-R and *g* loadings using principal axis factoring (PAF) for the WISC-III and the WAIS-III. Elliott (1990) reported loadings using maximum-likelihood estimation procedures from the DAS preschool battery (ages 4 to 5:11) and school-age battery (ages 6 to 17:11). McGrew (personal communication, November 21, 2003) reported *g* loadings using PCA for the WJ III for ages 4 to 5, 9 to 13, and 20 to 39. Kaufman and Kaufman (1994, 2004a) reported *g* loadings using PAF for the KABC-II (ages 7 to 18) and the KAIT (ages 20 to 34).⁵Subtests may measure more than one ability. ⁶Some evidence suggests that both constituent subtests in the composite measure the same *primary* narrow ability.

sonal communication, November 21, 2003; Sattler, 2001; Sattler & Ryan, 2001; Sattler & Saklofske, 2001). Although the published g loadings of subtests stem from large, nationally representative standardization samples from batteries including multiple reliable subtests measuring several broad abilities—which is ideal for measuring g (Jensen & Weng, 1994)— g loadings may vary somewhat as a function of the composition of the battery (Thorndike, 1987). To address this concern and to standardize all of the g loadings for subtests, principal component analysis was used to obtain g loadings for each of the six samples used in this study by selecting *all* cognitive ability subtests included in each sample and extracting a single principal component. Across the six samples, the number of subtests included in the analysis ranged from 24 to 34, and sample sizes for these analyses ranged from 76 to 144.

In Table 1, the published g loadings are represented on the left side of the sixth column from the left, and the g loadings conducted for this study are included on the right side of that column. In cases in which more than one data set contained subtests contributing to composites included in the study, their g loadings were averaged across data sets. Because of the notable variation in g loadings of subtests for the three age groups in which the WJ III was used, g loadings for subtests and composite g loadings for the WJ III composites are reported for each age group. Subtest g loadings of .70 or higher are considered *high*, those from .50 to .69 *medium*, and those below .50 *low* (McGrew & Flanagan, 1998; cf., Kaufman, 1994). Due to the absence of g loadings reported for composites, mean g loadings for composites were calculated by averaging the g loadings for subtests in each composite. The composite g loadings are reported in the far right side of Table 1.

Analyses

All exchangeability indices were computed using the total samples.⁶ Data screening procedures were conducted, and exchangeability was examined using several techniques. As evident in Table 2, analyses included calculating Pearson product-moment correlation co-

efficients (r) between each pair of broad ability composite scores. Due to many composite score distributions displaying restriction of range during preliminary analysis, Pearson correlations were corrected for restriction of range (Cohen, Cohen, West, & Aiken, 2003; Gulliksen, 1987). Next, the differences between the mean scores for each pair of composites were reported (see DM column in Table 2). These values stem from subtracting the mean of the composite stemming from the more recently normed test battery (e.g., the KABC-II) from the mean of the composite from the test battery normed earlier (e.g., the WISC-III).

To quantify the extent and magnitude of score differences between composites, *each individual's* norm-based composite score (i.e., a deviation IQ score) stemming from the more recently normed battery was subtracted from the corresponding composite score stemming from the test battery normed earlier. Because these difference values may be negative (if scores from the test battery normed later are higher) or positive (if scores from the test battery normed earlier are higher), with a relatively normal distribution, the mean difference would be 0 or near it because the negative and positive difference values tend to cancel each other out. To address this issue, Table 2 presents the mean absolute value of the difference between the composites (see MADI column). Thus, the average difference between scores—regardless of whether the difference is negative or positive—is presented.

Table 2 also presents the percentage of participants who displayed agreement, or non-significant differences, between composite scores. Two methods were used that stem from the distribution of the absolute value of the differences between scores. The first method stems from guidelines from McGrew and Flanagan (1998). Using this method, if the difference between composite scores for an individual—regardless of whether negative or positive—is less than or equal to 10 points, the composites were classified as agreeing for that individual.⁷ (See A10 column in Table 2.) Although this approach to agreement between composites is useful because it uses the same

Table 2
Exchangeability Statistics for Broad Ability Composites by Broad Ability

Composites Used in Each Comparison	S	N	U		DM	MADI	A10 ^a	ACI ^b	AD ^c	T ^c	P x T, E
			r	C							
<i>Crystallized Intelligence (Gc)</i>											
WPPSI-R Gc4, DAS Verbal, & WJ III Gc	1	130	—	—	—	—	—	—	.70	0%	30%
WPPSI-R Gc4 minus DAS Verbal	1	152	.81	.87	-.06	7.73	72%	86% (14.64)	.76	0%	24%
WPPSI-R Gc2:1 minus DAS Verbal	1	152	.65	.73	-1.32	9.97	60%	84% (16.40)	.62	0%	38%
WPPSI-R Gc2:2 minus DAS Verbal	1	152	.73	.78	.67	9.04	64%	84% (16.40)	.71	0%	30%
WPPSI-R Gc2:3 minus DAS Verbal	1	152	.73	.79	.57	8.94	65%	88% (16.40)	.70	0%	30%
WPPSI-R Gc2:4 ^d minus DAS Verbal	1	152	.76	.81	-.68	8.46	68%	87% (16.40)	.74	0%	26%
WPPSI-R Gc2:5 minus DAS Verbal	1	152	.78	.84	-.78	8.06	72%	86% (15.12)	.74	0%	26%
WPPSI-R Gc2:6 minus DAS Verbal	1	152	.81	.84	1.21	7.75	70%	92% (16.40)	.79	0%	21%
WPPSI-R Gc4 minus WJ III Gc	1	134	.68	.76	-1.57	7.32	76%	83% (12.61)	.67	1%	32%
WPPSI-R Gc2:1 minus WJ III Gc	1	134	.62	.70	-3.08	6.97	66%	81% (14.37)	.59	3%	38%
WPPSI-R Gc2:2 minus WJ III Gc	1	134	.57	.62	-.55	9.02	66%	82% (14.37)	.56	0%	44%
WPPSI-R Gc2:3 minus WJ III Gc	1	134	.63	.70	-.73	7.90	71%	85% (14.37)	.62	0%	38%
WPPSI-R Gc2:4 ^d minus WJ III Gc	1	134	.62	.68	-2.41	8.83	65%	82% (14.37)	.60	1%	38%
WPPSI-R Gc2:5 minus WJ III Gc	1	134	.68	.75	-2.60	7.63	73%	81% (13.10)	.67	2%	31%
WPPSI-R Gc2:6 minus WJ III Gc	1	134	.61	.65	-.06	8.80	67%	84% (14.37)	.61	0%	39%
DAS Verbal minus WJ III Gc	1	133	.68	.77	-1.80	9.15	63%	78% (15.12)	.65	1%	35%

(Table 2 continues)

(Table 2 continued)

Composites Used in Each Comparison	S	N	U		r	DM	MADI	A10 ^a	ACI ^b	AD ^c	T ^c	P x T, E
			C	C								
WISC-III VCI minus WJ III Gc	2	148	.79	.86	.97	6.85	76%	87% (12.12)	.77	0%	22%	
WISC-III Gc2:1 minus WJ III Gc	2	148	.75	.83	-1.21	7.47	78%	87% (13.49)	.74	0%	26%	
WISC-III Gc2:2 minus WJ III Gc	2	148	.73	.81	-2.19	7.98	66%	85% (14.27)	.70	1%	28%	
WISC-III Gc2:3 minus WJ III Gc	2	148	.78	.85	-.33	6.58	82%	89% (13.49)	.78	0%	22%	
WISC-III Gc2:4 ^d minus WJ III Gc	2	148	.70	.78	.39	7.67	74%	85% (14.99)	.69	0%	31%	
WISC-III Gc2:5 minus WJ III Gc	2	148	.74	.82	2.25	7.30	76%	91% (14.27)	.73	2%	25%	
WISC-III Gc2:6 minus WJ III Gc	2	148	.72	.80	1.27	7.30	72%	91% (14.64)	.71	0%	28%	
WISC-III VCI minus KABC-II Gc	4	116	.80	.82	2.31	7.71	73%	89% (13.06)	.79	1%	20%	
WISC-III Gc2:1 minus KABC-II Gc	4	116	.78	.80	2.14	7.31	75%	86% (14.43)	.77	1%	22%	
WISC-III Gc2:2 minus KABC-II Gc	4	116	.73	.76	.07	8.57	66%	83% (15.21)	.73	0%	27%	
WISC-III Gc2:3 minus KABC-II Gc	4	116	.83	.85	1.11	6.60	81%	91% (14.43)	.83	0%	17%	
WISC-III Gc2:4 ^d minus KABC-II Gc	4	116	.67	.70	1.73	9.72	57%	86% (15.92)	.67	1%	33%	
WISC-III Gc2:5 minus KABC-II Gc	4	116	.77	.79	2.77	7.67	72%	88% (15.21)	.75	2%	23%	
WISC-III Gc2:6 minus KABC-II Gc	4	116	.72	.75	.70	8.88	66%	86% (15.57)	.72	0%	28%	
WISC-III VCI minus KABC-II Gc2:1	4	116	.80	.85	2.78	7.65	71%	82% (13.06)	.76	2%	22%	
WISC-III Gc2:1 minus KABC-II Gc2:1	4	116	.78	.84	2.61	7.26	75%	86% (14.43)	.76	2%	23%	
WISC-III Gc2:2 minus KABC-II Gc2:1	4	116	.74	.81	.54	7.95	72%	91% (15.21)	.72	0%	28%	
WISC-III Gc2:3 minus KABC-II Gc2:1	4	116	.81	.86	1.57	6.79	78%	86% (14.43)	.80	1%	20%	

(Table 2 continues)

(Table 2 continued)

Composites Used in Each Comparison	S	N	U		r	DM	MADI	A10 ^a	ACI ^b	AD ^c	T ^c	P x T, E
			C	C								
WISC-III Gc2:4 ^d minus KABC-II Gc2:1	4	116	.68	.75	.75	2.20	8.84	66%	85% (15.92)	.67	1%	32%
WISC-III Gc2:5 minus KABC-II Gc2:1	4	116	.76	.82	.82	3.23	7.67	75%	90% (15.21)	.73	3%	25%
WISC-III Gc2:6 minus KABC-II Gc2:1	4	116	.72	.79	.79	1.16	8.19	73%	86% (15.57)	.70	0%	30%
WISC-III VCI minus KABC-II Gc2:2 ^d	4	116	.78	.83	.83	4.07	8.41	65%	80% (13.06)	.74	4%	22%
WISC-III Gc2:1 minus KABC-II Gc2:2 ^d	4	116	.76	.81	.81	3.90	8.04	76%	84% (14.43)	.73	4%	23%
WISC-III Gc2:2 minus KABC-II Gc2:2 ^d	4	116	.73	.79	.79	1.83	8.73	71%	87% (15.21)	.71	1%	29%
WISC-III Gc2:3 minus KABC-II Gc2:2 ^d	4	116	.79	.84	.84	2.87	7.48	78%	85% (14.43)	.77	2%	21%
WISC-III Gc2:4 ^d minus KABC-II Gc2:2 ^d	4	116	.67	.73	.73	3.49	9.61	64%	80% (15.92)	.64	3%	32%
WISC-III Gc2:5 minus KABC-II Gc2:2 ^d	4	116	.74	.80	.80	4.53	8.79	70%	86% (15.21)	.69	5%	26%
WISC-III Gc2:6 minus KABC-II Gc2:2 ^d	4	116	.71	.77	.77	2.46	8.66	69%	85% (15.57)	.69	1%	30%
WJ III Gc minus KABC-II Gc	5	83	.80	.83	.83	.00	6.77	81%	89% (13.06)	.80	0%	20%
WJ III Gc minus KABC-II Gc2:1	5	83	.80	.85	.85	-.47	6.66	80%	84% (13.06)	.79	0%	21%
WJ III Gc minus KABC-II Gc2:2 ^d	5	83	.79	.83	.83	1.34	6.77	75%	84% (13.06)	.79	0%	21%
<i>Visual Processing (Gv)</i>												
WPPSI-R Gv3, DAS Gv2, & WJ III Gv	1	138	—	—	—	—	—	—	—	.54	1%	45%
WPPSI-R Gv3 minus DAS Gv2	1	150	.57	.66	.66	-2.58	9.62	65%	87% (17.50)	.55	2%	42%
WPPSI-R Gv2:1 minus DAS Gv2	1	150	.57	.64	.64	-.39	8.86	67%	90% (18.78)	.57	0%	43%
WPPSI-R Gc2:2 minus DAS Gv2	1	150	.53	.64	.64	-3.85	9.53	65%	84% (17.50)	.50	5%	45%

(Table 2 continues)

(Table 2 continued)

Composites Used in Each Comparison	S	N	U		r	DM	MADI	A10 ^a	ACI ^b	AD ^c	T ^c	P x T, E
			C	r								
WPPSI-R Gc2:3 minus DAS Gv2	1	150	.50	.55	-.349	9.96	63%	89% (20.18)	.48	4%	48%	
WPPSI-R Gv3 minus WJ III Gv	1	144	.51	.62	-.18	8.77	61%	84% (17.50)	.51	0%	49%	
WPPSI-R Gv2:1 minus WJ III Gv	1	144	.49	.57	1.73	10.05	62%	83% (18.78)	.49	1%	51%	
WPPSI-R Gv2:2 minus WJ III Gv	1	144	.48	.59	-1.33	10.00	59%	84% (17.50)	.48	0%	52%	
WPPSI-R Gv2:3 minus WJ III Gv	1	144	.45	.50	-.93	10.70	58%	87% (20.18)	.45	0%	55%	
DAS Gv2 minus WJ III Gv	1	141	.55	.64	1.76	9.67	57%	86% (17.15)	.55	1%	45%	
WISC-III Gv2 minus WJ III Gv	2	148	.32	.36	1.88	12.56	49%	78% (20.60)	.32	0%	68%	
DAS Spatial minus WJ III Gv	3	121	.19	.21	12.93	17.63	41%	59% (18.34)	.14	26%	60%	
WISC-III Gv minus KABC-II Gv	4	116	.64	.67	-1.74	8.11	59%	85% (17.83)	.63	0%	37%	
WISC-III Gv minus KABC-II Gv2:1	4	116	.57	.63	.32	9.59	64%	91% (18.18)	.57	0%	43%	
WISC-III Gv minus KABC-II Gv2:2	4	115	.51	.64	1.22	9.22	72%	89% (20.33)	.49	0%	51%	
WISC-III Gv minus KABC-II Gv2:3	4	116	.63	.68	.58	8.47	72%	87% (17.09)	.62	0%	38%	
WISC-III Gv minus KABC-II Gv2:4	4	115	.56	.67	1.52	8.91	67%	89% (18.85)	.55	0%	44%	
WISC-III Gv minus KABC-II Gv2:5	4	115	.48	.59	2.61	10.35	58%	86% (19.47)	.47	2%	51%	
WJ III Gv minus KABC-II Gv	5	83	.54	.54	-3.34	11.51	59%	80% (19.92)	.50	2%	48%	
WJ III Gv minus KABC-II Gv2:1	5	83	.54	.54	-1.00	10.78	58%	87% (20.27)	.47	0%	53%	
WJ III Gv minus KABC-II Gv2:2	5	82	.52	.52	-.91	10.97	54%	90% (22.41)	.39	0%	61%	
WJ III Gv minus KABC-II Gv2:3	5	83	.49	.56	-.63	10.52	54%	86% (19.17)	.48	0%	52%	

(Table 2 continues)

(Table 2 continued)

Composites Used in Each Comparison	S	N	U		r	DM	MADI	A10 ^a	ACI ^b	AD ^c	T ^e	P x T, E
			U	C								
WJ III Gv minus KABC-II Gv2:4	5	82	.39	.50	.50	-.48	11.40	48%	84% (20.93)	.38	0%	62%
WJ III Gv minus KABC-II Gv2:5	5	82	.33	.43	.43	1.29	12.25	51%	83% (21.55)	.32	0%	68%
<i>Fluid Reasoning (Gf)</i>												
DAS Nonverbal minus WJ III Gf	3	128	.68	.71	.71	1.28	8.89	68%	80% (13.36)	.68	0%	32%
WJ III Gf minus KABC-II Gf ^d	5	83	.63	.63	.63	4.49	11.24	59%	65% (13.74)	.60	3%	36%
KAIT Gf2 minus WJ III Gf	6	80	.46	.60	.60	2.21	8.54	73%	73% (11.50)	.45	1%	54%
<i>Processing Speed (Gs)</i>												
WISC-III PSI minus WJ III Gs	2	146	.63	.66	.66	5.51	10.34	63%	82% (17.01)	.58	7%	35%
WAIS-III PSI minus WJ III Gs	6	146	.61	.69	.69	2.94	9.62	60%	75% (14.99)	.56	7%	37%

Note: S = sample number; U = uncorrected correlation; C = correlation corrected for restriction of range; DM = difference between group-based mean values for each pair of composites; MADI = mean of the absolute value of the differences between scores for individuals on each pair of composites; A10 = agreement 10; ACI = agreement confidence interval; AD = absolute unit dependability coefficient; T = percentage of variance attributed to the test battery; P x T, E = percentage of variance attributed to (a) the interaction between the individuals and the test battery and (b) random error; DAS = Differential Abilities Scales; KAIT = Kaufman Adolescent and Adult Intelligence Test; KABC-II = Kaufman Assessment Battery for Children—Second Edition; WAIS-III = Wechsler Adult Intelligence Scale—Third Edition; WISC-III = Wechsler Intelligence Scale for Children—Third Edition; WPPSI-R = Wechsler Preschool and Primary Scale of Intelligence, Revised; WJ III = Woodcock-Johnson III Tests of Cognitive Abilities; PSI = Processing Speed Index.

^aThe percentage of participants in the sample who demonstrated a difference of 10 or fewer points between scores on each pair of composites. ^bThe percentage of participants in the sample who demonstrated a difference of less than or equal to the sum of their respective 90% confidence interval values between scores on each pair of composites. ^cBecause negative variances are likely due to sampling error, they were set to 0 (Brennan, 2001; Cronbach et al., 1972). ^dSome evidence suggests that both constituent subtests in the composite measure the same primary narrow ability.

standard to compare all composite pairs, it does not consider the actual reliability of the composites, and it purports to represent a relatively small confidence interval (i.e., 68%). To address these concerns, a second analysis of agreement included calculation of 90% confidence intervals based on the reliability coefficients for composites presented in Table 1 (Charter & Feldt, 2000). If the difference between composite scores for an individual—regardless of sign—was less than or equal to the value associated with the sum of half of the 90% confidence interval for each composite, the composites were classified as agreeing for that individual (see ACI column in Table 2). For example, for the WPPSI-R *Gc4* minus DAS Verbal comparison, the confidence interval value for the WPPSI-R *Gc4* is 6.06 and the value for DAS Verbal is 8.58. Their sum is 14.64, and score differences less than or equal to the composites were classified as agreeing. To facilitate comparison between the two types of agreement described here, the specific values for the 90% confidence interval analysis are reported in parentheses in the ACI column in Table 2.

Generalizability theory was used to estimate the dependability of broad ability composite scores and to determine the amount of variance in obtained test scores associated with different test batteries and with error. These results are presented on the right side of Table 2. Variance components were obtained using the General Linear Model and Variance Components subprograms from SPSS 12.0. For each comparison, variance components were calculated that represented differences in individuals' measured ability that is systematically due to the test battery used (*T* in Table 2) and to error from the interaction between the individuals and the test battery as well as to random error (*P* x *T*, *E* in Table 2). These variance components were then used to calculate the absolute unit dependability coefficients.⁸ (See AD column in Table 2.) These coefficients are analogous to reliability coefficients and represent the extent to which a single composite score can be generalized to other composite scores measuring the same broad ability (Shavelson & Webb, 1991).

Results

Crystallized Intelligence (*Gc*)

As evident in Table 2, all correlation coefficients between pairs of *Gc* composite scores were statistically significant at $p < .001$. For the total sample, uncorrected correlations ranged from .57 to .83 ($M = .73$), and corrected correlations ranged from .62 to .87 ($M = .79$). The differences between the means for each pair of composites ranged from $-.3.08$ to 4.53. The average absolute value of the difference between *Gc* composite scores was about 8 points, with a range of averages from 6.6 to 10.0 points. Using the criterion of a difference greater than 10 points, an average of 71% of the participants from each sample demonstrated agreement between the two composites. When the actual confidence intervals were used, the average percentage of agreement was about 86%. This higher percentage of agreement stems from a wider range of true scores being accounted for by the 90% confidence interval based on actual reliability estimates.

As evident in the first row of Table 2, the absolute error unit dependability coefficient using *Gc* composite scores from three test batteries was .70 from Sample 1. Dependability coefficients for each possible pair of *Gc* composite scores tended to be lower, and none of the absolute unit dependability coefficients were above .80 ($M = .72$, range = .56 to .83). Many assert that scores used to make decisions about individuals should have reliability of at least .90 (e.g., Nunnally & Bernstein, 1994; Salvia & Ysseldyke, 2003). This same standard can be applied for generalizability coefficients, because in the case of this research, the issue in question is the reliability across measures of the same ability. A more liberal standard is .80. Review of variance components in Table 2 indicates that the influence of the test battery, per se, was negligible across all comparisons. In fact, it accounted for no more than 5% of the variance in any comparison, and in most cases it was 0. In contrast, the influence of systematic error stemming from the interaction between the individuals and the test battery and random

error was sizeable in all comparisons. The influence of these types of error accounted for an average of 28% of the variance across composite comparisons.

Visual Processing (*Gv*)

As evident in Table 2, all correlations between pairs of *Gv* composite scores were statistically significant. Correlations ranged from .19 to .64 ($M = .49$), and corrected correlations ranged from .21 to .68 ($M = .56$). The differences between the means for each pair of composites ranged from -3.85 to 12.93. The average absolute value of the difference between *Gv* scores was about 10 points, with a range of averages from 8.1 to 17.6 points. Using the criterion of a difference greater than 10 points, an average of 59% of the participants from each sample demonstrated agreement between the two composites. The range of agreement was 41% to 72%. When the actual confidence intervals were used, the average percentage of agreement was about 85%. Again, this notably higher percentage of agreement stems from a wider range of true scores being accounted for by use of the 90% confidence interval.

As evident from the first row in the *Gv* section of Table 2, the absolute error unit dependability coefficient stemming from analysis of three *Gv* composites from Sample 1 was .54. None of the dependability coefficients for pairs of *Gv* composite scores were above .80 ($M = .48$). In fact, five coefficients were less than .40. Review of variance components in Table 2 indicates that the influence of the test battery was negligible across all comparisons except one. In only one case was the percentage of variance due to test battery sizeable (DAS Spatial minus WJ III *Gv*). Again, in most cases it was 0. In contrast, the influence of systematic error stemming from the interaction between individuals and the test battery *and* random error was great in all comparisons. The influence of these types of error accounted for an average of more than 50% of the variance across composite comparisons. Clearly, it is error leading to the low generalizability coefficients for the *Gv* composites.

Fluid Reasoning (*Gf*)

All three correlation coefficients between *Gf* composite scores were statistically significant. Correlations ranged from .46 to .68 ($M = .59$), and corrected correlations were .60 to .71 ($M = .65$). The differences between the means for each pair of composites ranged from 1.28 to 4.49. The average absolute value of the difference between *Gf* composites was 9.6 points, with a range of averages from 8.5 to 11.2. Percentage agreement was relatively similar using the 10-point criterion ($M = 67%$, range of 59% to 73%) and the actual 90% confidence intervals ($M = 73%$, range from 65% to 80%). With one exception, more than one quarter of the participants in each sample demonstrated significant *differences* between the composites across the two methods of determining agreement. None of the absolute unit dependability coefficients was near .80 ($M = .58$, range = .45 to .68). Review of variance components indicates that the influence of the test battery was again very small across all comparisons. In contrast, the influence of systematic error stemming from the interaction between the individuals and the test battery *and* random error was sizeable in all comparisons. The influence of these types of error accounted for an average of more than 40% of the variance across composite comparisons.

Processing Speed (*Gs*)

Both correlation coefficients between pairs of *Gs* composite scores were statistically significant. Correlations were .61 to .63, and corrected correlations were .66 to .69. The differences between the means for each pair of composites ranged from 2.94 to 5.51. The average absolute value of the difference between *Gs* composites was 10 points, with a range of averages from 9.6 to 10.3. Using the criterion of a difference greater than 10 points, an average of 62% of the participants from each sample demonstrated agreement between the two composites. When the actual confidence intervals were used, the average percentage of agreement was about 78%, with a range from 75% to 82%. Neither of the absolute unit dependability coefficients was near .80 ($M = .57$).

Review of variance components indicates that the influence the test battery was the highest for the Processing Speed composites, on average across all ability comparisons, yet it was not sizeable. In both composite comparisons, the percentage of total variance accounted for by the influence of the test battery was only about 7%. Consistent with the findings for other broad ability composites, the influence of systematic error stemming from the interaction between the individuals and the test battery *and* random error was sizeable in all comparisons. The influence of these types of error accounted for an average of 36% of the variance across composite comparisons. It is notable that this percentage of variance is lower than those for the *Gv* and *Gf* composites.

Discussion

School psychologists and others engaged in psychological assessment may frequently assume that norm-based scores purported to measure the same ability should be relatively similar for individuals no matter what subtests or batteries were administered to obtain these scores. Information related to this assumption is important because consistency *across* measures of the same broad ability is needed in order to have confidence in interpreting normative or relative strengths or weaknesses in broad ability profiles. Using six samples of preschool children, school-age children, or adults who completed two or more intelligence test batteries, composites measuring the broad abilities Crystallized Intelligence, Visual Processing, Fluid Reasoning, and Processing Speed were compared. Results stemming from exchangeability analyses suggest that the CHC broad ability scores a person receives may vary greatly for some individuals and vary more so for some abilities than others. Visual Processing composites as a group appear to display notably low levels of exchangeability. An important question for practitioners and researchers is: What influences score exchangeability?

Improbable Influences

The statistics stemming from generalizability theory indicate (in all but a

handful of comparisons) that differences in obtained scores for composites are *not* due in any real sense to the characteristics of the batteries from which subtests were selected. In most all cases, the variance due to this set of influences was negligible, which is consistent with the generally small differences between composite means. Thus, when forming a broad ability composite, an examiner's choice of battery should not be expected to produce scores measuring broad abilities that are systematically higher or lower than those measuring the same abilities that stem from other batteries. Characteristics of the test battery, such as sampling of participants in the norming sample and the dates from which normative data were collected, do not appear to have consistent and systematic effects across individuals.

Although it may be assumed that all more recently normed tests produce lower scores for individuals than all tests normed earlier (Flynn, 1984), this finding was not uniform across composite comparisons. Of the composite comparisons, 44 (60%) yielded a positive difference between means, which would indicate that the test battery normed *earlier* produced a higher score on average. In contrast, 29 (40%) of the comparisons yielded a negative value for this difference, which indicates that the test normed *more recently* produced a higher score on average.⁹ It appears that expectations applied to IQs from subsequent revisions of one test battery (e.g., WISC-III to WISC-IV; Wechsler, 1991, 2003) should not be applied to related comparisons across test batteries—at least not at the broad ability composite level. However, additional research should evaluate the effect of cohort sampling on the exchangeability of ability scores (Rodgers, 1999; Sanborn, Truscott, Phelps, & McDougal, 2003).

Differences in the scaling of subtest scores and the formation of composite scores may also contribute to score differences for individuals. For example, score differences may have stemmed from the different scaling of the standardized scores from subtests of different test batteries. For example, each score increment for scaled scores ($M = 10$, $SD = 3$) rises and falls in units representing 1/3 of a

standard deviation (i.e., 5 standard score points), whereas *T* scores and deviation IQ scores have increments of 1/10 and one 1/15 of a standard deviation, respectively. However, these effects do not appear to be systematic in producing higher or lower scores across individuals.

Probable Influences

The study's results indicate that most of the variability in the obtained composite scores that was *not* due to measurement of true ability stemmed from (a) random error and (b) *interactions* between the examinee and characteristics of the test battery and its constituents.

Random error. The extent to which random error should be expected to influence a composite score is represented by the composite reliability coefficient and the resulting standard error of measurement and confidence interval. Only 21 broad ability composites (53%) included in the analysis demonstrated reliability coefficients of .90 or greater, which can be considered excellent or high (Bracken, 1987; Cicchetti, 1994; McGrew & Flanagan, 1998). Table 1 reveals that slightly less than half of the composites (47%) demonstrated reliability below that level. For all of these composites, a 90% confidence interval spans more than 16 deviation IQ points, or more than half a standard deviation (8 points) above and below the obtained score. Less reliable composites lead to less confidence that the obtained norm-based score will well represent the examinee's true ability on an absolute level. As a result, somewhat large score differences between composite scores with low reliability can be expected to be nonsignificant and commonplace.

Examinee–task interactions. In general, the interactions between the examinees and the test batteries can be grouped into three categories: (a) the interaction between the examinee's ability level and *score characteristics*, (b) the interaction between characteristics of the examinee and the *temporal aspects* of the testing sessions, and (c) the interaction between characteristics of the examinee and the *requirements of the assessment tasks* (i.e., subtests; see Bracken, 1988).

The interaction between the examinee's ability level and subtest or composite score characteristics is a probable culprit leading to less exchangeability (Anastasi & Urbina, 1997; Salvia & Ysseldyke, 2003). For example, if the range of items on a subtest is not sufficient to tap into the ability of those with very low or very high ability, the individual's true ability will be inaccurately represented by the obtained score. For example, reviews of published evaluations of subtest floor and ceiling violations and review of actual norms tables and output from computer-based scoring programs indicated that a few of the subtests included in the composites in this study would be expected to demonstrate inadequate floors or ceilings (Kaufman & Kaufman, 2004a; McGrew & Flanagan, 1998; Schrank & Woodcock, 2001; Tusing, Maricle, & Ford, 2003). There were a few subtest floor or ceiling violations with the DAS (Elliott, 1990) with school-age children (Sample 3) and with the KAIT (Kaufman & Kaufman, 1993) with the adults (Sample 6). However, a sizeable number of scores from the youngest children with low ability in Sample 1 (i.e., those ages 3 to 4) would be expected to be affected notably by inadequate floors. Similarly, for participants with very low or very high ability across samples, test battery-based composites would be expected to produce scores that are more deviant from the normative mean than the derived composites, which stem from averaging subtest scores.

The interaction between characteristics of the examinee and the temporal aspects of the testing sessions also may contribute error to the measurement of true ability in a number of documented ways (see Saklofske & Schwean Kowalchuk, 1994; Zhu & Tulskey, 2000). These interactions may be reflected in effects attributed to practice, test order, or fatigue on some subtests and not others, the amount of time between test sessions, and varying motivation or propensity to guess at different times during the assessment session. It is likely that the counterbalancing used in the collection of the data from the six samples led to this interaction and not to a systematic influence leading to

higher or lower scores on one battery versus another.

Finally, the interaction between characteristics of examinee and the requirements of the assessment tasks (i.e., subtests) probably affect obtained scores—even when obvious influences like sensory acuity deficits and English language proficiency are considered (see Dean & Woodcock, 1999; McGrew & Flanagan, 1998). On one level, subtests may be selected to measure broad abilities that tap into exceptionally well-developed or poorly developed skills of the individual (e.g., CHC narrow abilities). Subtests may also tap into construct-irrelevant abilities or other influences rather than the broad ability being measured, such as speed on a task designed to measure Visual Processing, fine motor skills on a task designed to measure Processing Speed, or vocabulary knowledge on tasks measuring Visual Processing (Bracken, 1986, 2000). Some tasks may also be more familiar or appealing to some examinees in a manner affecting their performance. The choice of item content, per se, may differentially affect some examinees (e.g., numbers versus letters with Short-Term Memory subtests), and response requirements (e.g., use of manipulatives) may have differential effects for others (Kyllonen, 1995).

In summary, this study's analysis of the influences on score exchangeability indicates that examiners can randomly select a composite within a broad ability area, and selection of the composite will not contribute independent error to the score that results. However, the composite's selection may interact with characteristics of an examinee to misrepresent the examinee's broad ability. The reliability of the composite must also be considered.

Influence of *g* and Similarity in Qualitative Characteristics

It is logical that the more a composite score measures *g*—what is common across all subtests in a battery—the greater its exchangeability with composites from other batteries (Jensen, 1998; Spearman, 1927). The *g* loadings for composites presented in Table 1 and exchangeability indexes in Table 2 seem to indicate this pattern. In all analyses except the

one using the 90% confidence interval, which considers actual reliability of the composite, comparisons between composites with the highest *g* loadings seem to demonstrate the highest level of exchangeability and vice versa. Perhaps composites with higher *g* loadings may simply be more reliable, and reliability has the greatest effect on exchangeability. In contrast to the focus on *g*, it is possible that shared qualitative characteristics of the subtests contributing to composites affect exchangeability. The relations between qualitative characteristics, such as similarity in the narrow abilities measured by the composites, similarity in cultural and linguistic loadings of subtests across composites (Ortiz, 2002; Ortiz & Ochoa, 2005), similarity in elementary information processes measured by subtests across composites (Carroll, 1976; Floyd, 2005), should also be examined for their effects on composite exchangeability. Additional research is needed to examine these issues concurrent with consideration of composite reliability.

Limitations and Caveats

First, it could be argued that all six samples of participants used in this study are not representative of the United States population. However, they were drawn from multiple, largely independent sites, and the samples were generally large. The samples may also be faulted by some for including restricted ranges of talent, including young children who may have displayed problem behaviors that undermine reliable assessment, and including adults with diagnostic classifications. The concern about range restriction is partially addressed with the use of correlation corrections, and analysis of subsamples not reported in this article revealed that exchangeability indexes for individuals with diagnostic classifications in Sample 6 did not differ notably from those from individuals with no such classifications. Furthermore, results appear similar across samples. Second, although the data used in these studies match in many ways what one would see in practice—ability composite scores extracted from test batteries administered in a standard fashion—the data sets are archival. Thus, undocumented aspects of the

data collection process may have affected score exchangeability. It is clear that more research should be done to examine these influences with additional samples and other batteries measuring CHC abilities.

Third, although steps were taken to obtain published *g* loadings for subtests and to compute them based on joint exploratory factor analysis, the derivations of the *g* loadings for composites may be considered by some to be a weak aspect of this research. It seems important for researchers and test authors to calculate and report *g* loadings and specificity estimates for composites stemming from test batteries. Although these properties are often reported for subtests, they are very rarely reported for composites (cf. Elliott, 1990, pp. 190-192). Fourth, some may assert that reporting agreement by examining score differences greater than the sum of half of the 90% confidence interval value for one composite plus the half of that value from the other composite is not optimal. Perhaps the standard error of the difference should have been used to determine agreement between composites, or perhaps estimated true scores and the standard error of estimate values could have been used. Based on a review of these issues (e.g., Charter & Feldt, 2000) and calculation of these alternate statistics, the approach employed was used for several reasons, including (a) the values used were the largest (i.e., most facilitative of agreement) of the three alternatives and (b) the approach probably best represents the steps that test users take to examine the significance of score differences.

Finally, this research was not a direct examination of any particular interpretive framework for CHC-based assessment. With the onslaught of guidelines for CHC-based interpretation of broad abilities composites by test authors and other experts, it is unlikely that all of the steps deemed necessary to interpretation in recent publications were represented fully in these analyses. Therefore, additional research may be needed to examine these issues.

Implications for Practice

The findings from the exchangeability analyses are probably no surprise to many

school psychologists and others engaged in psychological assessment because they recognize error in measurement. At best, this research provides evidence supporting best practices in cognitive ability assessment. One common recommendation is to consider reliability (and unreliability) during score interpretation (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Excellent practice-oriented textbooks, such as Sattler (2001) and Kamphaus (2001), provide clear descriptions of issues related to reliability and recommendations for use of confidence intervals. Based on these recommendations and this research, those engaged in testing should use large confidence intervals, such as the 90% and 95% confidence intervals, to represent the range of true scores around obtained scores (or estimated true scores). The standard practice of reporting the nominal labels for multiple score ranges (e.g., Low Average to Average) should also be encouraged when the confidence interval range indicates that it is necessary. In addition, it is important to consider the actual confidence intervals for composites rather than considering rules-of-thumb (e.g., ± 5 points). Similarly, test users should (a) be wary of or avoid interpreting composites with low reliability or (b) expect that large differences between composites with low reliability and other composites may be expected due to unreliability in measurement.

School psychologists and others engaged in psychological assessment can continue to engage in practices in order to control for or reduce the negative influences of idiosyncratic responses to test stimuli, task requirements, or response requirements. They first should strive to reduce error by design. It is possible that an understanding of both (a) the conditions during completion of cognitive tasks (i.e., subtests) under which an examinee responds well and (b) conditions in which the examinee responds poorly may be valuable for informing instruction or interventions. In that vein, some knowledgeable test users do *not* want duplicate measures of a broad ability because they desire measures sensitive to evoking those differences

in performance. However, it seems that the paramount goal of ability testing should be to measure accurately the examinee's ability without evidence of the undermining influences of interactions among the task, setting, situation, and examinee characteristics. Thus, based on known characteristics of intelligence test batteries and known characteristics of the examinee, test users can choose the test battery that will ensure sound measurement of CHC broad abilities. These steps are logical and enhanced by knowing well the test battery, its scores, and their properties. Examiners also should continue to observe influences on examinee responding, such as level of motivation, fatigue, and strategy use, and note if these influences undermine measurement of the targeted ability. These observations may be enhanced by using rating scales for test session behaviors, such as those included on many test records. Examiners may also benefit from systematically examining patterns of errors (see Kaufman & Kaufman, 2004b).

In essence, test users should continue to devote their energies and insights to well-integrated assessments and accurate and direct measurement of abilities. According to Meyer et al. (1997), "A psychological test is a dumb tool, and the worth of the tool cannot be separated from the sophistication of the clinician who draws inferences from it and then communicates with patients and other professionals" (p. 153). Whether engaged in ability, curriculum-based, or behavioral assessment or other assessment activities, well-informed school psychologists and other test users can continue to demonstrate their expertise in psychological assessment in a manner that leads to the improvement of the lives of children, adults, and their families. Greater knowledge about the potential influences on score exchangeability can contribute to this expertise.

Footnotes

¹ The value representing the difference between the mean of one measure and the mean of the another, $M_x - M_y = Z$, is equivalent to the value representing the mean of the differences between the same two measures across participants, $\Sigma(X - Y)/N$.

² In this article, the word *subtests* is used consistently to describe individual cognitive ability tasks.

³ Participants in Sample 3 and Sample 6 did not complete the WJ III General Information subtest; therefore, comparisons including the WJ III *Gc* composite, which includes the General Information test, were not included for these samples.

⁴ Four composites were cautiously constructed for which there was some evidence that both constituent subtests measured the same *primary* narrow ability. These composites were included because there was also evidence that at least one of the subtests that shared a primary narrow ability also measured at least one *other* narrow ability. In fact, for three of the four composites, both subtests contributing to them had been judged to measure more than one narrow ability within the same broad ability area. Each composite is designated in Table 1 and Table 2.

⁵ One exception to this rule was the exclusion of composites from Sample 1 that included the Matching Letter-Like Forms subtest from the DAS (Elliott, 1990). We judged that too few children in Sample 1 completed this subtest and that its extant measurement characteristics were too poor to include in our analysis.

⁶ In addition to the results presented here, results from subsamples were also produced. Consistent with guidelines presented by McGrew and Flanagan (1998), subsamples of participants were selected whose subtests *within each composite* demonstrated confidence bands (± 7) that overlapped or touched (i.e., they differed no more than 14 deviation IQ score points). Because results were so remarkably similar to those from the analysis of the total sample and due to space limitations, these results are not reported here. These results can be obtained from the first author or at <http://www.psyc.memphis.edu/faculty/floyd/floydExCHC1.htm>

⁷ McGrew and Flanagan (1998) reported that "the SE_M estimates for all cross-battery composite averages . . . were set at ± 5 . . . (representing the average [median] *Gf-Gc* composite . . . SE_M scores across all ranges in the WJ-R norm sample)" (pp. 390, 398). Note that the SE_M value is equivalent to the 68% confidence interval.

⁸ The formula for the absolute unit dependency coefficient is the following:

$$\frac{\sigma_p^2}{\frac{\sigma_p^2 + \sigma_t^2}{1} + \frac{\sigma_{pxt,e}^2}{1}}$$

⁹ Although the mean and median values for the positive differences were somewhat larger in magnitude than the negative differences, the discrepancies between these values were less than one deviation IQ score point.

References

- Alfonso, V. C., Flanagan, D. P., & Radwan, S. (2005). The impact of the Cattell–Horn–Carroll theory on test development and interpretation of cognitive and academic abilities. In D. P. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed., pp. 185–202). New York: Guilford Press.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.
- Bergman, L. R., & Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Development and Psychopathology*, 9, 291–319.
- Bracken, B. A. (1986). Incidence of basic concepts in the directions of five commonly used American tests of intelligence. *School Psychology International*, 7, 1–10.
- Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment*, 5, 313–326.
- Bracken, B. A. (1988). Ten psychometric reasons why similar tests produce dissimilar results. *Journal of School Psychology*, 26, 155–166.
- Bracken, B. A. (2000). Maximizing construct relevant assessment: The optimal preschool testing situation. In B. A. Bracken (Ed.), *The psychoeducational assessment of preschool children* (3rd ed., pp. 33–44). Boston: Allyn & Bacon.
- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.
- Cairns, R. B., Bergman, L. R., & Kagan, J. (Eds.). (1998). *Methods and models for studying the individual*. Thousand Oaks, CA: Sage.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new structure of intellect. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 27–56). Hillsdale, NJ: Erlbaum.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York: Cambridge University.
- Carroll, J. B. (2003). The higher stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5–21). New York: Pergamon.
- Charter, R. A., & Feldt, L. S. (2000). The relationship between two methods of evaluating an examinee's difference scores. *Journal of Psychoeducational Assessment*, 18, 125–142.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6, 284–290.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.
- Dean, R. S., & Woodcock, R. W. (1999). *The WJ–R and Batería–R in neuropsychological assessment* (Research Report No. 3). Itasca, IL: Riverside.
- Elliott, C. D. (1990). *Differential Ability Scales*. San Antonio, TX: Psychological Corporation.
- Evans, J. J., Floyd, R. G., McGrew, K. S., & Leforgee, M. H. (2002). The relations between measures of Cattell–Horn–Carroll (CHC) cognitive abilities and reading achievement during childhood and adolescence. *School Psychology Review*, 31, 246–262.
- Flanagan, D. P., & Kaufman, A. S. (2004). *Essentials of WISC–IV assessment*. New York: Wiley.
- Flanagan, D. P., & McGrew, K. S. (1997). A cross-battery approach to assessing and interpreting cognitive abilities: Narrowing the gap between practice and cognitive science. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 314–325). New York: Guilford Press.
- Flanagan, D. P., McGrew, K. S., & Ortiz, S. (2000). *The Wechsler Intelligence Scales and Gf–Gc theory: A contemporary approach to interpretation*. Needham Heights, MA: Allyn & Bacon.
- Flanagan, D. P., & Ortiz, S. (2001). *Essentials of cross-battery assessment*. New York: Wiley.
- Floyd, R. G. (2005). Information-processing approaches to interpretation of contemporary intellectual assessment instruments. In D. P. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed., pp. 203–233). New York: Guilford Press.
- Floyd, R. G., Bergeron, R., & Alfonso, V. C. (2005). *Cognitive ability profiles of children with reading comprehension difficulties*. Manuscript submitted for publication.
- Floyd, R. G., Clark, M. H., & Shadish, W. R. (2005). *The exchangeability of general ability composites*. Manuscript submitted for publication.
- Floyd, R. G., Evans, J. J., & McGrew, K. S. (2003). Relations between measures of Cattell–Horn–Carroll (CHC) abilities and mathematics achievement across the school-age years. *Psychology in the Schools*, 60, 155–171.
- Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, 95, 29–51.
- Glutting, J. J., Watkins, M. W., & Youngstrom, E. A. (2003). Multifactorial and cross-battery ability assessments: Are they worth the effort? In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence*,

- aptitude, and achievement* (2nd ed., pp. 343–374). New York: Guilford Press.
- Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale, NJ: Erlbaum.
- Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, & R. W. Woodcock, *WJ-R technical manual* (pp. 197–232). Itasca, IL: Riverside Publishing.
- Horn, J. L., & Blankson, N. (2005). Foundation for better understanding cognitive abilities. In D. P. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed., pp. 41–68). New York: Guilford Press.
- Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.
- Jensen, A. R., & Weng, L.-J. (1994). What is good g? *Intelligence*, 18, 231–258.
- Kamphaus, R. W. (2001). *Clinical assessment of child and adolescent intelligence* (2nd ed.). Boston: Allyn & Bacon.
- Kamphaus, R. W., Winsor, A. P., Rowe, E. W., & Kim, F. (2005). A history of intelligence test interpretation. In D. P. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed., pp. 23–38). New York: Guilford Press.
- Kaufman, A. S. (1994). *Intelligent testing with the WISC-III*. New York: John Wiley & Sons.
- Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children*. Circle Pines, MN: American Guidance Service.
- Kaufman, A. S., & Kaufman, N. L. (1993). *The Kaufman Adolescent and Adult Intelligence Test*. Circle Pines, MN: American Guidance Service.
- Kaufman, A. S., & Kaufman, N. L. (2004a). *Kaufman Assessment Battery for Children, Second Edition manual*. Circle Pines, MN: American Guidance Service.
- Kaufman, A. S., & Kaufman, N. L. (2004b). *Kaufman Test of Educational Achievement, Second Edition manual*. Circle Pines, MN: American Guidance Service.
- Kyllonen, P. C. (1995). CAM: A theoretical framework for cognitive abilities measurement. In D. Detterman (Ed.), *Current topics in human intelligence: Vol. 4. Theories of intelligence* (pp. 307–359). Norwood, NJ: Ablex.
- McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment*, 8, 289–302.
- McGrew, K. S. (2005). The Cattell–Horn–Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 136–177). New York: Guilford Press.
- McGrew, K. S., & Flanagan, D. P. (1998). *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment*. Boston: Allyn & Bacon.
- McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock–Johnson III technical manual*. Itasca, IL: Riverside Publishing.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education.
- Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, 56, 128–165.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Ortiz, S. (2002). Best practices in nondiscriminatory assessment. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 1321–1336). Washington, DC: National Association of School Psychologists.
- Ortiz, S., & Ochoa, S. H. (2005). Advances in cognitive assessment of culturally and linguistically diverse individuals. In D. P. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed., pp. 234–250). New York: Guilford Press.
- Phelps, L., McGrew, K. S., Knopik, S. N., & Ford, L. A. (2005). The general (g), broad, and narrow CHC stratum characteristics of the WJ III and WISC-III tests: A confirmatory cross-battery investigation. *School Psychology Quarterly*, 20, 66–88.
- Proctor, B., Floyd, R. G., & Shaver, R. B. (2005). CHC broad cognitive ability profiles of low math achievers. *Psychology in the Schools*, 42, 1–12.
- Rizza, M. G., McIntosh, D. E., & McCunn, A. (2001). Profile analysis of the Woodcock–Johnson III Tests of Cognitive Abilities with gifted students. *Psychology in the Schools*, 38, 447–455.
- Rodgers, J. L. (1999). A critique of the Flynn effect: Massive IQ gains, methodological artifacts, or both? *Intelligence*, 26, 337–356.
- Roid, G. H. (2003). *Stanford–Binet Intelligence Scale, Fifth Edition*. Itasca, IL: Riverside Publishing.
- Saklofske, D. H., & Schwan Kowalchuk, V. L. (1994). Influences on testing and test results. In M. Zeidner & R. Most (Eds.), *Psychological testing: An inside view* (pp. 89–118). Palo Alto, CA: Consulting Psychologists.
- Salvia, J., & Ysseldyke, J. (2003). *Assessment in special and remedial education* (8th ed.). Boston: Houghton Mifflin.
- Sanborn, K. J., Truscott, S. D., Phelps, L., & McDougal, J. L. (2003). Does the Flynn effect differ by IQ level in samples classified as learning disabled? *Journal of Psychoeducational Assessment*, 21, 145–159.
- Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th. ed.). San Diego: Author.
- Sattler, J. M., & Ryan, J. J. (2001). Wechsler Adult Intelligence Scale–III: Description. In J. M. Sattler, *Assessment of children: Cognitive applications* (4th. ed., pp. 375–413). San Diego: Author.
- Sattler, J. M., & Saklofske, D. H. (2001). Wechsler Intelligence Scale for Children–III: Description. In J. M. Sattler, *Assessment of children: Cognitive applications* (4th. ed., pp. 220–265). San Diego: Author.
- Schrank, F. A., & Woodcock, R. W. (2001). *WJ III Compuscore and Profiles Program* [Computer software]. Itasca, IL: Riverside Publishing.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, NY: Sage.
- Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York: Macmillan.
- Thorndike, R. L. (1987). Stability of factor loadings. *Personality and Individual Differences*, 8, 585–586.
- Tusing, M. B., Maricle, D. E., & Ford, L. A. (2003). Assessment with the Woodcock–Johnson III and young

- children. In F. A. Schrank & D. P. Flanagan (Eds.), *WJ III clinical use and interpretation* (pp. 244–283). New York: Academic Press.
- Wechsler, D. (1989). Wechsler Preschool and Primary Scale of Intelligence–Revised. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (1991). The Wechsler Intelligence Scale for Children–Third Edition. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (1997). Wechsler Adult Intelligence Scale–Third Edition. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (2003). The Wechsler Intelligence Scale for Children–Fourth Edition. San Antonio, TX: Psychological Corporation.
- Woodcock, R. W. (1990). The theoretical foundations of the WJ-R measures of cognitive ability. *Journal of Psychoeducational Assessment, 8*, 231–258.
- Woodcock, R. W., & Johnson, M. B. (1989). Woodcock-Johnson Psycho-Educational Battery–Revised. Allen, TX: DLM Teaching Resources.
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson III Tests of Cognitive Abilities. Itasca, IL: Riverside Publishing.
- Zhu, J., & Tulsy, D. S. (2000). Co-norming of the WAIS-III and WMS-III: Is there a test-order effect on IQ and memory scores? *The Clinical Neuropsychologist, 14*, 461–467.

Randy G. Floyd is an Assistant Professor of Psychology at The University of Memphis. He received his doctoral degree in School Psychology from Indiana State University. His research interests include assessment of cognitive abilities, identification of reading and mathematics aptitudes, and improving behavioral assessment methods.

Renee Bergeron received her MS in School Psychology at The University of Memphis in 2003, and she is a doctoral candidate in School Psychology at The University of Memphis. Currently, she is completing an internship at the Pearl City complex in Hawaii. Her research interests are in the areas of cognitive ability and behavioral assessment.

Allison C. McCormack received her MA in School Psychology at Middle Tennessee State University in 2002, and she is currently a doctoral candidate in School Psychology at The University of Memphis. Her research interests are in the areas of cognitive ability and neuropsychological assessment.

Janice L. Anderson is currently a student in the Community Agency Counseling program at The University of Memphis.

Gabrielle L. Hargrove-Owens received her BA at The University of Memphis in 2004. She is currently employed at the Robertson County, Tennessee Department of Children's Services. Her interests focus on the provision of psychological services in the schools.