Professional Psychology: Research and Practice
2008, Vol. 39, No. 4, 414–423

0735-7028/08/$12.00 DOI: 10.1037/0735-7028.39.4.414

The Exchangeability of IQs: Implications for Professional Psychology

Randy G. Floyd and M. H. Clark
University of Memphis

William R. Shadish
University of California, Merced

IQs are important measures in the practice of psychology. Psychologists may frequently expect that IQs from different test batteries are reasonably exchangeable as measures of general intelligence. Results presented in this article provide evidence that different test batteries produce less similar IQs for samples of school-age children and undergraduate students than may have been expected. In fact, psychologists can anticipate that 1 in 4 individuals taking an intelligence test battery will receive an IQ more than 10 points higher or lower when taking another battery. Resulting suggestions for practice include carefully choosing batteries that provide representative sampling of specific abilities, differential weighting, or both; attending to unreliability in measurement; closely monitoring behaviors that undermine assessment of general intelligence; and considering the benefits of obtaining multiple IQs when such scores are used to make high-stakes diagnostic or eligibility decisions.

Keywords: intelligence, IQ, exchangeability, score differences, cognitive abilities

Supplemental materials: http://dx.doi.org/10.1037/0735-7028.39.4.414.supp

Decisions by psychologists engaged in psychological and educational assessment are influenced by the IQs[1] yielded by the intelligence test batteries they administer. They place great weight on these scores when making decisions about their clients when disorders, such as mental retardation, reading disorder, mathematics disorder, and disorder of written expression, as well as intellectual giftedness, are in question. It is likely that most psychologists engaged in assessment practices possess several intelligence test batteries as well as the requisite competence to administer and interpret all of them. Psychologists are certainly bombarded with newly published test batteries and revisions of established ones that are purported to provide more valid measures of intelligence than their predecessors. However, in practice, it is probable that most psychologists typically use only one intelligence test battery for the majority of assessment-related referrals. They likely use one battery with the assumption that other well-designed and validated batteries would yield similar IQs for their clients. This article examines whether this assumption is warranted and offers recommendations for ensuring valid assessment of IQs.

Ubiquity and Importance of IQs

Intelligence test batteries are critical tools in psychology and education. Recent survey results have indicated that intelligence test batteries are often considered by psychology training directors as being essential for practice (e.g., Belter & Piotrowski, 2001) and that they are among the most frequently used assessment instruments in clinics, schools, and forensic settings (Camara, Nathan, & Puente, 2000; Rabin, Barr, & Burton, 2005; Ryba, Cooper, & Zapf, 2003; Wilson & Reschly, 1996). Clinicians are apparently well informed in their use of these test batteries because the batteries produce scores that hold great meaning. Across a large body of research, IQs have been shown to predict socially important variables, including academic attainment (e.g., achievement test scores, grades, and years of schooling), job performance, occupa-

RANDY G. FLOYD received his PhD in school psychology from Indiana State University. He is an associate professor in the Department of Psychology at the University of Memphis. His research interests include assessment of cognitive abilities, evaluation of behavioral assessment methods, identification of reading and mathematics aptitudes, and validation of general outcome measures tapping early numeracy skills.

M. H. CLARK received her PhD in quantitative psychology at the University of Memphis. She is an assistant professor at Southern Illinois University, Carbondale. Her research interests include selection bias in nonrandomized experiments, assessing bias in randomized studies with differential attrition, statistical adjustments to correct for those biases, propensity scores, and methods for creating and controlling parameters in computer simulations.

WILLIAM R. SHADISH received his PhD in clinical psychology from Purdue University. He is professor and founding faculty, University of California, Merced. His research interests include experimental and quasi-experimental design, the empirical study of methodological issues, the methodology and practice of meta-analysis, and evaluation theory.

WE THANK the Woodcock–Muñoz 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, M. B. Tusing, Susan League, LeAdelle Phelps, David McIntosh, Mardis Dunham, Noel Gregg, and Cheri Hoy. We thank Renee Bergeron, Jill Bose-Deakins, Bruce Bracken, Mark Daniel, Ron Dumont, Tom Fagan, Noel Gregg, Allison Margulies, Kevin McGrew, LeAdelle Phelps, and Fredrick Schrank for their comments on drafts of the manuscript. We are also appreciative of information from Larry Evans and Richard Shavelson. Portions of this research were presented at the annual meetings of the American Evaluation Association (2002), the Society for Multivariate Experimental Psychology (2003), and the National Association of School Psychologists (2004).

CORRESPONDENCE CONCERNING THIS ARTICLE should be addressed to Randy G. Floyd, Department of Psychology, University of Memphis, Memphis, TN 38152. E-mail: rgfloyd@memphis.edu

[1] We use *IQ* to describe scores, stemming from aggregation of multiple measures of cognitive abilities, that represent individual differences in general intelligence.

tional and social status, and income (Brody, 1997; Gottfredson, 2003; Jensen, 1998; Schmidt, 2002). Consequently, an IQ is typically required for the identification of individuals with mental retardation, learning disabilities, or intellectual giftedness (American Psychiatric Association [APA], 2000; Individuals With Disabilities Education Act, 1997; National Research Council, 2002). For example, recent summaries of the practices for identifying children in need of special education because of mental retardation and learning disabilities revealed that 48 of the United States require consideration of IQ for identification of mental retardation (Bergeron, Floyd, & Shands, in press) and that 48 states required consideration of an IQ–achievement discrepancy for identification of a learning disability (Reschly & Hosp, 2004).

In many respects, the characteristics of IQs have been widely studied and are clearly understood. In addition to the research examining the predictive properties of IQs, much research has used factor-analytic methods to examine the latent variables that influence performance on intelligence tests[2] that contribute to IQs. Many have asserted that the single, higher order latent variable general intelligence (or the *g* factor) is typically the largest and most important influence on all such tests (e.g., Carroll, 1993; Jensen, 1998). Another body of research has examined scores from successive revisions and renorming of intelligence test batteries across decades and revealed rising levels of intelligence in the population as a whole (i.e., the Flynn effect; Flynn, 1987; Neisser, 1998). As a result, it is expected that an individual displaying the same level of intelligence would receive a lower IQ on a test battery using more recent norms than that individual would on a version of the same battery normed earlier.

Exchangeability

We argue in this article that the practices and policies associated with psychological assessment should be informed regarding a key practical assumption about IQs—their exchangeability. In this context, *exchangeability* refers to the assumption that the IQ a person receives will be reasonably constant no matter which intelligence test battery is used. Comprehensive listings of published assessment instruments include a multitude of individually administered intelligence test batteries (Hammill, Brown, & Bryant, 1992; Kamphaus, Petoskey, & Rowe, 2000; Murphy, Spies, & Plake, 2006), and a recent search of the Buros Institute of Mental Measurement Web site (http://www.unl.edu/buros/bimm/index.html) yielded 250 published instruments measuring "intelligence and general aptitude." Although there is a consensus that the rising levels of intelligence reflected in more recent norm samples affect the IQ an individual receives on successive revisions of the same test battery, little is known about the exchangeability of IQs across different test batteries (cf. Flynn, 2006).

Psychologists may believe that exchangeability is hardly possible on the basis of a surface-level review of the assorted materials included in different intelligence test batteries, their varying administration formats, and their diverse task requirements. For example, some test batteries require the examinee to manipulate objects, such as blocks, whereas others use no such manipulatives. Some batteries require the examinee to provide detailed oral responses, whereas some others require no oral response from the examinee. Some batteries include tests that appear to vary widely in the measurement of abilities more specific than general intelligence, whereas others appear to measure a more narrow range. If IQs are conglomerates or mixture measures of specific abilities that have distinct developmental trajectories, genetic and neurological origins, and patterns of external relations, IQ stemming from different sets of specific ability measures would not be expected to produce exchangeable IQs (e.g., Horn & Blankson, 2005). However, exchangeability is thought to be plausible because of the principle of aggregation (Rushton, Brainerd, & Pressley, 1983). That is, influences associated with individual tests in a battery are averaged out when multiple test scores are aggregated into an IQ. As a result, only a single ability, general intelligence, is thought to remain as the systematic source of variance (Jensen, 1998; Spearman, 1927).

Decisions about those taking intelligence tests are made individual by individual, and although IQs are not the only piece of information used in those decisions, they frequently play a key role. Often, laws, regulations, and policies specify a particular IQ as a threshold for some categorization of an individual without specifying which intelligence test battery should be used to yield that score. If so, a person's IQ should be reasonably similar over different batteries. Decisions about whether a child is diagnosed with mental retardation, or whether a prisoner is executed, should not vary depending on which test battery they took (Kanaya, Scullin, & Ceci, 2003; Keyes, Edwards, & Perske, 2002; Koocher, 2003). Just as consumers expect that thermometers sold by different manufacturers should yield similar temperatures (when using the same scale), we expect that duplicate measures of a psychological construct should yield similar results for the same individual. We investigated this expectation of exchangeability of IQs to inform the practice of psychology.

The Investigation

We obtained data from six samples previously described in published test manuals (Kaufman & Kaufman, 2004; McGrew & Woodcock, 2001) and in subsequently published journal articles (Floyd, Bergeron, McCormack, Anderson, & Hargrove-Owens, 2005; Phelps, McGrew, Knopik, & Ford, 2005). Participants from each of these samples completed, in counterbalanced order, at least two intelligence test batteries that yielded IQs. Participants in five of the six samples were children within the age range of 8 to 16. Participants in the sixth sample were undergraduate students.[3]

Table 1 presents characteristics of the intelligence test batteries and the IQs used in this study. The table includes the years in which normative data were collected, the number of tests contributing to the IQ, the specific latent factors measured by tests, the median internal consistency coefficients across select ages. and the actual test–retest reliability coefficients for the IQs. Note that the Woodcock–Johnson III Tests of Cognitive Abilities (WJ III; Woodcock, McGrew, & Mather, 2001) produces two IQs that stem

[2] We use *tests*, rather than *subtests*, to describe individual cognitive ability tasks. We use *intelligence test battery* to describe collections of these tests that produce IQs.

[3] Information about the samples is available online in our supplemental materials.

Table 1
Measurement Characteristics of Test Batteries and Associated IQs

Test battery and IQ	Years norm data collected	No. of tests included in IQ	Specific latent factors measured by tests	Median internal consistency coefficient[a]	Test-retest reliability coefficient (retest interval)[b]	Median *g* loading of tests (range of *g* loadings)[c]
DAS GCA (Elliott, 1990)	1987-1989	6	Verbal, Nonverbal, Spatial	.95	.93 (2–6 weeks)	.70 (.63–.76)
KAIT CIQ (Kaufman & Kaufman, 1993)	1988-1991	6	Fluid, Crystallized	.97	.91 (6–99 days)	.78 (.68–.84)
KABC-II FCI (Kaufman & Kaufman, 2004)	2001-2003	10	Short-Term Memory, Visual Processing, Long-Term Storage and Retrieval, Fluid Reasoning, Crystallized Ability	.97	.92 (12–56 days)	.62 (.45–.78)
WAIS-III FSIQ (Wechsler, 1997)	1995	11	Verbal Comprehension, Perceptual Organization, Working Memory, Processing Speed	.98	.91 (2–12 weeks)	.72 (.57–.83)
WISC-IV FSIQ (Wechsler, 2003)	1998-2003	10	Verbal Comprehension, Perceptual Organization, Working Memory, Processing Speed	.97	.89 (13–63 days)	.64 (.82–.48)
WISC-III FSIQ (Wechsler, 1991)	1989	10	Verbal Comprehension, Perceptual Organization, Freedom From Distractibility, Processing Speed	.97	.95 (12–63 days)	.72 (.44–.79)
WJ III GIA-Standard (Woodcock et al., 2001)	1996-1999	7	Comprehension-Knowledge, Long-Term Retrieval, Visual-Spatial Thinking, Fluid Reasoning, Short-Term Memory, Auditory Processing, Processing Speed	.96 (ages 7–12) .98 (ages 18–55)		.61 (.51–.79) (ages 9–13) .71 (.58–.81) (ages 20–39)
WJ III GIA-Extended (Woodcock et al., 2001)	1996-1999	14	Comprehension-Knowledge, Long-Term Retrieval, Visual-Spatial Thinking, Fluid Reasoning, Short-Term Memory, Auditory Processing, Processing Speed	.97		.57 (.30–.79)

Note. Information about norming, specific factors measured, and reliability was obtained from the technical manuals supporting each test battery: DAS (Elliott, 1990); KAIT (Kaufman & Kaufman, 1993); KABC-II (Kaufman & Kaufman, 2004); WAIS-III (Wechsler, 1997); WISC-IV (Wechsler, 2003); WISC-III (Wechsler, 1991); and WJ III (McGrew & Woodcock, 2001). DAS GCA = Differential Abilities Scales General Conceptual Ability; KAIT CIQ = Kaufman Adolescent and Adult Intelligence Test Composite Intelligence Quotient; KABC-II FCI = Kaufman Assessment Battery for Children, Second Edition, Fluid–Crystallized Index; WAIS-III FSIQ = Wechsler Adult Intelligence Scale—Third Edition Full Scale IQ; WISC-IV FSIQ = Wechsler Intelligence Scale for Children—Fourth Edition Full Scale IQ; WISC-III FSIQ = Wechsler Intelligence Scale for Children—Third Edition Full Scale IQ; WJ III GIA-Standard = Woodcock–Johnson III Tests of Cognitive Abilities General Intellectual Ability–Standard; WJ III GIA–Extended = Woodcock–Johnson III Tests of Cognitive Abilities General Intellectual Ability–Extended.

[a] Unless otherwise noted, the reported coefficients are values for the approximate age ranges included in this study.

[b] The reported coefficients are uncorrected values for the approximate age ranges included in this study.

[c] All test *g* loadings, except those for the DAS, WISC-IV, and WJ III tests, resulted from principal-axis factoring (Kaufman & Kaufman, 1993, 2004; Sattler, 2001). Test *g* loadings for the DAS resulted from maximum likelihood estimation using a model specifying a first-order *g* factor, test *g* loadings for the WISC-IV tests resulted from maximum likelihood estimation using a model specifying a second-order *g* factor (Keith, Fine, Taub, Reynolds, & Kranzler, 2006), and test *g* loadings for the WJ III tests resulted from principal-components analysis (Kevin S. McGrew, personal communication, November 21, 2003). When possible, the *g* loadings are values for the approximate age ranges included in this study.

from seven of the same tests.[4] To provide an index of the degree to which each IQ measures general intelligence, Table 1 also includes the median and range of *g* loadings for the tests included in the IQs. The *g* loadings represent the relations between (a) individual test scores included in a battery and (b) the single factor from factor analysis that represents general intelligence (Jensen, 1998). Squared *g* loadings yield the proportion of variance in tests attributable to this factor. Test *g* loadings of .70 or higher are considered good, those from .50 to .69 are considered fair, and

[4] Because our goal was to compare IQs across intelligence test batteries, we did not analyze the exchangeability of the WJ III GIA–Standard and the WJ III GIA–Extended.

Table 2
Select Exchangeability Statistics for IQ Comparisons

IQs used in each comparison	Diff norm dates	Sample				Exchangeability statistics			
		No.	*n*	Age range	Pearson *r*	MD	MAD	A10[a]	ACI[b] (critical value)
WISC-III FSIQ and WJ III GIA-Standard	10	1	148	8–12	.72	3.63	7.78	70.3	66.2 (9.90)
WISC-III FSIQ and WJ III GIA-Extended	10	1	146	8–12	.76	2.03	6.80	82.2	77.6 (9.24)
DAS GCA and WJ III GIA-Standard	12	2	120	8–12	.76	2.56	7.88	71.7	71.7 (10.48)
WJ III GIA-Standard and KABC-II FCI	7	3	83	8–12	.79	−0.05	7.42	74.7	69.9 (9.24)
WISC-III FSIQ and KABC-II FCI	14	4	116	8–12	.77	1.46	7.53	72.4	70.7 (9.82)
WISC-IV FSIQ and KABC-II FCI[c]	5	5	29	8–12	.93	−0.76	5.10	93.0	75.9 (8.57)
WISC-IV FSIQ and KABC-II FCI[c]	5	5	27	13–16	.87	−3.85	6.96	70.4	66.7 (8.57)
WAIS-III FSIQ and WJ III GIA-Standard	4	6	148	18–53	.69	9.39	10.70	55.4	35.8 (7.79)
KAIT CIQ and WAIS-III FSIQ	7	6	99	18–53	.75	−5.12	7.08	74.7	63.6 (7.00)
KAIT CIQ and WJ III GIA-Standard	11	6	147	18–53	.73	4.50	7.36	75.5	57.8 (7.79)

Note. Values in parentheses represent the specific values for the 90% confidence interval analysis. Diff norm dates = difference in years between beginning of norming of test battery normed first to end of norming for test battery normed second; No. = sample number; MD = mean of the differences between IQs; MAD = mean of the absolute value of the differences between IQs; WISC-III FSIQ = Wechsler Intelligence Scale for Children—Third Edition Full Scale IQ; WJ III GIA–Standard = Woodcock–Johnson III Tests of Cognitive Abilities General Intellectual Ability–Standard; WJ III GIA–Extended = Woodcock–Johnson III Tests of Cognitive Abilities General Intellectual Ability–Extended; DAS GCA = Differential Abilities Scales General Conceptual Ability; KABC-II FCI = Kaufman Assessment Battery for Children, Second Edition Fluid–Crystallized Index; WISC-IV FSIQ = Wechsler Intelligence Scale for Children—Fourth Edition Full Scale IQ; WAIS-III FSIQ = Wechsler Adult Intelligence Scale—Third Edition Full Scale IQ; KAIT CIS = Kaufman Adolescent and Adult Intelligence Test Composite Intelligence Quotient.
[a] The percentage of participants in the sample who demonstrated a difference of 10 or fewer points between each pair of IQs.
[b] The percentage of participants in the sample who demonstrated a difference of less than or equal to the sum of half their respective 90% confidence interval values between each pair of IQs.
[c] Children ages 8 to 12 complete 10 KABC-II tests, including the Triangles test. However, children ages 13 to 16 complete 9 of the same tests, but not Triangles. Instead, they complete Block Counting. Because of this reason and because the size of both subsamples are similar, the total sample was separated into two subsamples.

those below .50 are considered poor (Kaufman, 1994). Of the IQs listed in Table 1, only the WJ III General Intellectual Ability (GIA)–Standard and the WJ III GIA–Extended stem from aggregation of test scores that are differentially weighted according to their *g* loadings (McGrew & Woodcock, 2001). All other IQs stem from aggregation of equal-weighted test scores.

For each data set, Pearson product–moment correlation coefficients were computed for each pair of IQs and reported in Table 2. Not surprisingly, all correlations were statistically significant at $p < .001$. Correlations ranged from .69 to .93, and the average correlation was .78 (*Mdn* = .76). All correlations were in the moderate to very strong range.[5]

Table 2 also includes two sets of results examining differences between IQs, per se.[6] The differences between the mean scores for each pair of IQs were calculated and reported as mean differences in Table 2. These values were obtained by subtracting the mean of the IQ from the more recently normed test battery (e.g., the WJ III) from the mean of the IQ from the test battery normed earlier (e.g., the Wechsler Intelligence Scale for Children—Third Edition, or WISC-III).[7] The differences between the means for each pair of IQs ranged from −.5.12 to 9.39. To quantify the extent and magnitude of score differences between IQs for each individual, the absolute value of the difference between IQs for each participant was obtained. The means of these difference scores are reported in Table 2. The mean absolute value of the difference between IQs was about 7 points, with a range of values from 5.1 to 10.7 points.

Table 2 also presents the percentage of participants who displayed agreement, or nonsignificant differences, between IQs. Two methods were used that drew from the distribution of the absolute values of the differences between IQs. Using the first method, if the absolute difference between IQs for an individual was less than or equal to 10 points, the IQs were classified as agreeing for that individual (see the A10 column in Table 2). Using the 10-point criterion for determining agreement between IQs is useful because it uses the same standard across IQ comparisons. In addition, because many guidelines for diagnosis and identification of mental retardation recommend consideration of about 5 points above and below the obtained IQ (APA, 2000; Bergeron et al., in press), this standard likely reflects a common rule of thumb used by psychologists (Groth-Marnat, 2003). However, use of the 10-point criterion for evaluation of agreement does not allow for consideration of the actual reliability of the IQs. To address this limitation, a second analysis of agreement used the 90% confidence intervals plotted around obtained IQs (Charter & Feldt,

[5] We recognize that there is no standard rule of thumb for providing nominal labels for *r* values. We drew from the following general labels: *negligible,* .00 to .19; *weak,* .20 to .39; *moderate,* .40 to .69; *strong,* .70 to .89; and *very strong,* .90 to 1.0.

[6] Additional results of these analyses (i.e., standard deviations, skewness, and kurtosis values) are available online in our supplemental materials.

[7] The value representing the difference between the mean of one measure and the mean of another ($M_x - M_y = Z$) is mathematically equivalent to the mean of the sum of differences between the same two measures for individual participants [$\Sigma(X - Y)/N$)].

Table 3
Dependability Coefficients and Variance Components for IQ Comparison

IQs used in each comparison	Sample		Generalizability theory analysis		
	No.	*n*	AD	*T*[a] (%)	*P* × *T*, *E* (%)
Dependability over all IQs: WAIS-III FSIQ, KAIT CIQ, and WJ III GIA–Standard	6	96	.53	22	25
Dependability between pairs of IQs					
WISC-III FSIQ and WJ III GIA–Standard	1	148	.68	4	27
WISC-III FSIQ and WJ III GIA–Extended	1	146	.75	1	24
DAS GCA and WJ III GIA–Standard	2	120	.75	1	22
WJ III GIA–Standard and KABC-II FCI	3	83	.78	0	22
WISC-III FSIQ and KABC-II FCI	4	116	.76	0	23
WISC-IV FSIQ and KABC-II FCI	5	29	.93	0	7
WISC-IV FSIQ and KABC-II FCI	5	27	.85	2	13
WAIS-III FSIQ and WJ III GIA–Standard	6	148	.51	26	24
KAIT CIQ and WAIS-III FSIQ	6	99	.66	11	22
KAIT CIQ and WJ III GIA–Standard	6	147	.67	8	25

Note: No. = sample number; AD = absolute error unit dependability coefficient; *T* = percentage of variance attributed to the test battery, *P* × *T*, *E* = percentage of variance attributed to (a) the interaction between the individuals and the test battery and (b) random error; WAIS-III FSIQ = Wechsler Adult Intelligence Scale—Third Edition Full Scale IQ; KAIT CIQ = Kaufman Adolescent and Adult Intelligence Test Composite Intelligence Quotient; WJ III GIA–Standard = Woodcock–Johnson III Tests of Cognitive Abilities General Intellectual Ability–Standard; WISC-III FSIQ = Wechsler Intelligence Scale for Children—Third Edition Full Scale IQ; WJ III GIA–Extended = Woodcock–Johnson III Tests of Cognitive Abilities General Intellectual Ability–Extended; DAS GCA = Differential Abilities Scales General Conceptual Ability; KABC-II FCI = Kaufman Assessment Battery for Children, Second Edition Fluid–Crystallized Index; WISC-IV FSIQ = Wechsler Intelligence Scale for Children—Fourth Edition Full Scale IQ.
[a] Negative variances were set to 0 (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972)

2000). The confidence interval was calculated for this study on the basis of the internal consistency reliability coefficients for IQs presented in Table 1. If the absolute difference between IQs for an individual was less than or equal to the value associated with the sum of half of the 90% confidence interval for each IQ, the IQs were classified as agreeing for that individual (see the ACI column in Table 2).[8] For example, the 90% confidence interval for the WISC-III Full Scale IQ (FSIQ) was calculated to be less than or equal to 4.95 points, the 90% confidence interval for the WJ III GIA-Standard was calculated to be less than or equal to 4.95 points, and their sum was 9.90 points. Therefore, individuals with IQ differences that were less than or equal to 9.90 points were classified as having IQs that agreed. To facilitate comparison between the two types of agreement analysis, the specific values for the 90% confidence interval analysis are reported in parentheses in the ACI column in Table 2.

Using the 10-point criterion, an average of 74% of the participants from each sample demonstrated agreement between the two IQs. When the actual confidence intervals were used, the average percentage of agreement was about 66%. Using the actual confidence intervals reduced the percentage of agreement because the internal consistency reliability estimates for the IQs were very high (≥.95), thereby making the range of the 90% confidence intervals for each IQ more narrow than 10 points.

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, we used generalizability theory to measure the extent to which a single IQ can be generalized to other IQs and to provide information about the influences that have the largest effects on IQs, such as the influence stemming from the tests and test batteries themselves (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). Table 3 presents the resulting absolute error unit dependability coefficients (see the AD column) and variance components. An absolute error unit dependability coefficient treats both discrepant rank orderings of the same individual over IQs and discrepant mean IQs as error. The resulting coefficient measures the likelihood that the average IQ a person receives across *x* different IQs will be exactly the same as his or her population score, where *x* is the number of IQs used in computing the coefficient. Decisions about people in clinical practice are rarely based on the average of five or even two IQs; rather, they are usually based on just one. This dependability coefficient estimates the generalizability of the IQ based on just one test battery.

We presented dependability coefficients two ways in Table 3: (a) using three IQs obtained from Sample 6 and (b) using each pair of IQs from each sample. Results from examination of the exchangeability of the three IQs from Sample 6—the WAIS-III FSIQ, the Kaufman Adolescent and Adult Intelligence Test Composite Intelligence Quotient (KAIT CIQ), and WJ III GIA-Standard—revealed a dependability coefficient of only .53. For the remainder of the pairwise IQ comparisons across samples, the dependability coefficients were somewhat higher than the comparison using three batteries, but they were consistently lower than the

[8] The percentage of participants obtaining an IQ significantly higher or lower than the other IQ on the basis of the 90% confidence interval criteria is available online in our supplemental materials.

.90 standard ($M = .73$, $Mdn = .75$) often used when evaluating the reliability of scores for decisions about individuals in clinical practice (Nunnally & Bernstein, 1994; Salvia & Ysseldyke, 2003; Sternberg, 1994). Only one comparison meets or exceeds the .90 standard. Perhaps this standard is too high for dependability coefficients, so a standard of .80 may be more appropriate (R. J. Shavelson, personal communication, April 5, 2007). Only two additional IQ comparisons demonstrated a dependability coefficient at or above this level, and both IQ comparisons stemmed from consideration of the same two IQs (Wechsler Intelligence Scale for Children—Fourth Edition [WISC-IV FSIQ] and Kaufman Assessment Battery for Children, Second Edition Fluid–Crystallized Index [KABC-II FCI]).

Because the results of the exchangeability analyses suggest that the IQ a person receives may depend heavily on which test battery is used, we examined reasons for these differences. For each comparison, variance components were calculated that represented the difference in IQs that is systematically due to the test battery used (T in Table 3) and the difference that is due to error from the interaction between the individuals and the test battery as well as to random (and otherwise unspecified) error ($P \times T$, E in Table 3). Review of variance components indicates that the influence of the test battery per se was negligible across comparisons from Samples 1 through 5. In fact, it accounted for no more than 4% of the variance in any of these comparisons. In contrast, results from Sample 6 revealed that a sizable percentage of the variance was due to differences stemming from the test batteries themselves. In fact, the variance attributed to the test battery was more than 25% of the total variance in the WAIS-III FSIQ to WJ III GIA-Standard comparison. Results from all six samples indicated that the influence of error stemming from the interaction between the individuals and the test battery and random error was sizable. The influence of these types of error ranged from 7% to 27% and accounted for an average of 21% of the variance across IQ comparisons. Thus, most of the variability in IQs that is not due to true ability stems not from the test battery per se but from unsystematic and idiosyncratic responses to test stimuli, task requirements, or response requirements.

Implications for Professional Psychology

To inform best practices in the use and interpretations of IQs, we sought to examine the exchangeability assumption with IQs from some of the most well-developed and recently published intelligence test batteries. These IQs have demonstrated exceptionally high estimates of reliability as well as strong evidence of validity, and they are arguably some of the most psychometrically sound measures available to psychologists. On the basis of these positive judgments, measurement of general intelligence through IQs may appear to be misleadingly straightforward in the practice of psychology (Groth-Marnat, 2003).

Should Exchangeability of IQs Be Expected?

On one hand, psychologists should know that our results indicate that IQs from different test batteries demonstrate reasonably high exchangeability. They are all highly correlated with each other, and most tend to produce scores that are generally within the same range for the majority of individuals. In fact, across most exchangeability indexes, IQs appear to be more exchangeable than measures said to represent abilities more specific than general intelligence (cf. Floyd et al., 2005). For some psychologists, finding greater exchangeability of IQs may be surprising because IQs are yielded from a variety of tests that may vary greatly in their use of test stimuli, their administration formats, and their task requirements, whereas measures of more specific abilities appear more similar across these characteristics. The greater exchangeability of IQs is probably a result of greater reliability and the aggregation of multiple test scores and "averaging out" of more specific abilities so that general intelligence remains as the primary source of variance.

On the other hand, psychologists should know that IQs across intelligence test batteries may not be reasonably constant. For example, on the basis of our results, we can anticipate that more than one in four individuals taking an intelligence test battery would receive an IQ more than 10 points higher or lower when taking another test battery. In addition, considering that most expert opinion is that reliability estimates for important measures in psychology be no less than .80, our results indicate that only 2 out of 10 pairwise IQ comparisons meet or exceed this standard (Shavelson & Webb, 1991). On the basis of these results, we can conclude that different intelligence test batteries do tend to produce (for a number of reasons) very discrepant scores for at least some individuals.

Systematic effects on exchangeability. Although discrepant IQs can be anticipated for at least some individuals, our results from samples of school-age children indicate that these discrepancies are not typically due in any real sense to the global characteristics of the intelligence test batteries. In most cases, our results revealed that the variance resulting from this set of influences was negligible. Thus, in most cases, an examiner's choice of intelligence test battery should not be expected to produce IQs that are systematically higher or lower across school-age children than other batteries. Characteristics of the test battery, such as the demographic and ability-level representativeness of participants in the norming sample and the dates from which normative data were collected, should not be expected to have consistent and systematic effects on IQ for children such as those from our samples.

Our results indicate that, although the Flynn effect may influence test scores by producing, on average, lower IQs on more recently normed intelligence test batteries, these effects are relatively small when compared with other effects on test scores when the test batteries have been normed about a decade or less apart. Results from our analyses that best supported evidence of the Flynn effect (Flynn, 1987, 2006), our WISC-III FSIQ and WJ III GIA–Standard comparison, indicated a mean difference in IQs of 3.6 points that may have resulted because of an increase in this population's intelligence during the 10 years between the beginning of norming for the WISC-III and the end of norming for the WJ III. However, to put this reasonably small mean difference between IQs in perspective, about 30% of the IQs from this sample were different by more than 10 IQ points (see the MD and A10 columns in Table 2).

It is also apparent that the Flynn effect did not appear consistently across our IQ comparisons. Across the seven pairwise IQ comparisons stemming from samples of school-age children, only four demonstrated mean differences consistent with the trend expected as a result of the Flynn effect. Another comparison

yielded a mean difference nearing zero, and the other two comparisons yielded mean differences indicating that the more recently normed battery yielded higher IQs. The mean difference values, the typical indicators of the Flynn effect, were also inconsistent when variables in our comparisons were held constant. For example, in our two WISC-IV FSIQ and KABC-II FCI comparisons, which differed primarily according to the age of the sample (ages 8–12 vs. ages 13–16), the sizes of the mean differences were notably different (−0.76 and −3.85). Thus, although IQs were obtained from the same general norm samples for each test battery, the mean differences between IQs were notably different for the two groups, and these values are in the opposite direction of those expected because of the Flynn effect.

In contrast to the findings from the five samples of school-age children, differences in IQs from the sample of undergraduate students seem to be attributable to the differences in the global characteristics of the test batteries. In fact, when IQs from the three batteries we included are considered in concert, more than 20% of the total variance is attributable to the influence of the test batteries. When pairwise comparisons were considered, the WAIS-III produced IQs that were notably higher, on average, than the WJ III (about 9 points higher) and the KAIT CIQ (about 5 points higher), whereas the KAIT CIQ produced, on average, IQs that were more than 4 points higher than the WJ III. These findings are consistent with the assertions of Flynn (2006) that the WAIS-III has norms that are atypical and possibly substandard. Thus, although the WAIS-III was normed up to 7 years after the KAIT CIQ, it produced notably higher IQ scores on average. Furthermore, the mean IQ from the WAIS-III was more than 9 points higher than the WJ III, which was normed only approximately 4 years later. It may be, as Flynn has asserted, that obtaining adequate norming samples for adults who display the full range of individual differences in ability is challenging and rarely achieved. As a result, differences in the norming samples of adults may produce IQs in this population that are not exchangeable.

Unsystematic and interactive effects on exchangeability. Psychologists using intelligence test batteries should be aware that most of the variability in IQs that is not due to true ability stems from two general classes of influences represented in our results. The first class of influences includes what is considered random error. Thus, examinee behaviors, such as failure to retrieve known answers and use of detrimental strategies, as well as broad sampling of content and operations to which examinees may not have been exposed, may lead them to miss test items that are within their ability range. On the other hand, other examinee behaviors, such as guessing or use of facilitative strategies, as well as past serendipitous exposure to content and operations like those required for items may lead them to answer correctly items that are above their ability level. The effects of these types of unsystematic influences are represented by internal consistency reliability coefficients, and they are accounted for using the standard error of measurement and confidence intervals plotted around IQs. The second class of influences includes the interactions between the examinees and the global characteristics of the test battery. These interactions can be grouped into three categories: (a) interactions between characteristics of the examinee and the temporal aspects of the testing sessions, (b) interactions between the examinee's ability level and score characteristics, and (c) interactions between characteristics of the examinee and the requirements of the tests.

Interactions between characteristics of the examinee and the temporal aspects of the testing sessions may be reflected in effects attributed to practice, order of test administration, fatigue on some tests and not others, the amount of time between test sessions, and varying motivation or propensity to guess at different times during the assessment session. Participants in all samples included in this investigation completed a minimum of 13 tests to obtain at least one pair of IQs, and some completed more than 40 tests. Consistent with testing in schools, clinics, and other settings, most participants likely completed the testing during multiple sessions. Although all test batteries were administered in counterbalanced order, counterbalancing still leaves the possibility that these effects will vary depending on which test batteries are administered and the order in which they are administered.[9]

Although it is not apparent from our results because of the samples we included, interactions between (a) the examinee's ability level and (b) the characteristics of the IQs and their tests are probable culprits leading to less exchangeability in the practice of psychology. For example, if the range of items on a test is not sufficient to tap into the ability of those with very low or very high ability (i.e., there are inadequate floors and ceilings), the individual's true ability will be inaccurately represented by the obtained IQ. Similarly, perhaps inadequate exchangeability surfaces because IQs differ substantially in their measurement of general intelligence. At present, the precision of IQs in measuring general intelligence is typically extrapolated from the *g* loadings of tests contributing to the IQs. As is evident in Table 1, somewhat dissimilar medians and ranges of *g* loadings are evident for these tests. However, the utility of this comparison is lessened because the WJ III GIA–Standard and WJ III GIA–Extended stem from differential weighting of tests according to their *g* loadings, whereas the other IQs do not. Despite this issue, the varying precision at which each IQ measures general intelligence (vs. more specific abilities) likely contributes to the variability in scores for individuals (Bergeron & Floyd, 2006; Spitz, 1988).

Finally, interactions between characteristics of individual examinees and the requirements of the assessment tasks (i.e., tests) may affect IQs when influences not associated with general intelligence are tapped. For example, problems with fine motor skills, sensory acuity, or English language proficiency may unduly affect performance on some tests for some individuals so that the targeted ability is not measured well. Such effects may also be apparent on tests that are familiar or appealing to some individuals. In contrast to systematic effects attributed to the test battery when most examinees possess characteristics that likely affect IQs (e.g., high crystallized intelligence), these types of interactions reflect idiosyncratic responses by individuals to the tests in the battery.

Recommendations for Psychologists

Psychologists engaged in frequent assessment know well that the assessment process, which may include consideration of IQs, requires careful selection of assessment methods, integration of

[9] Analysis of IQs from Samples 3, 4, and 5 revealed that, in five of six analyses, there were no significant differences in the IQs from those test batteries administered first and those administered second. Results are available online in our supplemental materials.

different sources of information, and a good deal of reflection. On the basis of our findings, we offer the following recommendations for consideration during this process. First, at a minimum, psychologists should ensure that an IQ in question has strong reliability and a strong network of validity evidence supporting its use for the specific purpose for which it will be used (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Most IQs meet these criteria. Once an IQ has been yielded through testing, psychologists should typically ground their description of that IQ and what it measures by explicitly referencing the intelligence test battery from which it came.

Second, on the basis of our array of exchangeability analyses across IQs, it appears that psychologists should not expect the IQ of each individual (especially children) taking an intelligence test battery to be affected unilaterally and primarily by differences in the ability sampling reflected in its norms (i.e., the Flynn effect). Thus, these effects on IQs—often quoted to produce IQs about 3 points lower per decade between norming of test batteries—appear not to be uniform because of other, perhaps more powerful effects. Despite this finding, because evidence for the Flynn effect has been repeatedly demonstrated for identical and related measures of mental ability (e.g., Flynn, 1987; Neisser, 1998), psychologists should strive to administer recently normed intelligence test batteries. However, psychologists should consider perhaps more influential effects on IQs that are at play in the practice of psychology. These effects are those considered unsystematic and interactive.

Third, because our results indicate that much of the variability across IQs that is not due to measurement of true ability is due to random error, psychologists should consider the unreliability of IQs during interpretation by using 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). On the basis of research indicating that less than half of the United States require those involved in determining children's eligibility for mental retardation to attend to the unreliability of IQs (Bergeron et al., in press), policymakers at national, state, and local levels should ensure that test users consider confidence intervals around IQs. For example, if any part of the confidence interval reaches a recognized cut score, such as an IQ of 70 for mental retardation, that single diagnostic or eligibility criterion could be said to be met. Perhaps psychologists should routinely calculate and report, on the basis of an obtained IQ and its range of true scores, the probability of an IQ reaching such well-established IQ thresholds. In addition, the standard practice of reporting the nominal labels for multiple score ranges for IQs (e.g., low average to average) should also be encouraged when the confidence interval range spans more than one normative level (Groth-Marnat, 2003; Sattler, 2001). Recognizing this uncertainty in measurement of general intelligence will increase confidence that the IQ in question will be exchangeable with IQs from other test batteries.

Fourth, psychologists should continue to engage in practices to control for or to reduce the negative influences of idiosyncratic responses to test stimuli, task requirements, or response requirements. Knowing these results, psychologists should first strive to reduce these effects by design. It is possible that an understanding of both (a) the conditions during completion of cognitive tasks 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 psychologists do not want exchangeable IQs because they desire tasks sensitive to evoking those differences in performance. However, if the purpose of assessment is to determine the likely academic and occupational outcomes of those served by psychologists, which is likely the most valid interpretation of IQs, it seems that the paramount goal of intelligence testing should be to measure accurately the examinee's general intelligence without evidence of the undermining influences of the interactions among the task, setting, situation, and examinee characteristics (Floyd et al., 2005).

Through careful selection of the intelligence test battery and its tests, the examiner can avoid potentially undermining effects, such as those resulting, at least in part, from inadequate floors, inadequate ceilings, sensory acuity requirements of items, and task requirements for motor responses. Thus, on the basis of known characteristics of intelligence test batteries and known characteristics of the examinee, psychologists should choose the test battery and its tests that will ensure sound measurement of general intelligence via IQs. With more and more intelligence test batteries providing several alternate measures of general intelligence stemming from a greater number of tests, fewer tests, or alternate groups of tests[10] as well as the continued option of substituting tests, psychologists are probably better equipped than ever to produce valid IQs.

Examiners should also continue to observe influences on examinee responding during testing, such as level of motivation, fatigue, and strategy use, and note whether these influences undermine measurement of general intelligence (Meyer et al., 2001). These observations may be enhanced by using rating scales for global evaluation of the effects of test-session behaviors (e.g., Glutting & Oakland, 1993; McConaughy & Achenbach, 2004). Because these often deleterious but powerful influences on cognitive performance occur most apparently at the level of the individual test, psychologists may be well advised to consider the effects of these influences test by test, in addition to considering them globally at the completion of the testing session. Psychologists may also be well advised to describe these influences and their effects prominently in their psychological assessment reports to ensure that the inferences drawn from the testing represent the examinee's general intelligence well. They may also report, in addition, alternate but equally reliable and valid IQs from the same battery that exclude test scores that may have been unduly affected by these influences.

Fifth, when general intelligence is in question, psychologists may be well advised to choose an intelligence test battery that provides the most comprehensive and representative sampling of ability measures (Jensen, 1998). On review of Table 1, one test battery appears to measure only two specific latent abilities, and only two test batteries appear to measure five or more specific latent abilities. In addition to comprehensive and representative ability sampling, psychologists should also consider the usefulness of selecting IQs that stem from (a) weighting of their tests (e.g., Woodcock et al., 2001) versus using test batteries in which all tests (even those with low reliability and poor *g* loadings) contribute equally to the IQ or (b) multiple tests with uniformly good *g*

[10] Additional information about the relations between comprehensive IQs and abbreviated IQs is available online in our supplemental materials.

loadings. Representative sampling might reduce the likelihood that IQs will be unduly affected by redundancy in or absence of measurement of any particular specific ability, and using IQs stemming from weighting or from multiple, highly *g*-loaded tests appears to reduce the influence of specific variance from each test on IQs.

Sixth, psychologists could continue to use and to refine interpretive strategies designed to judge the influence of specific abilities (i.e., specific variance from test scores) on IQs. These specific abilities produce construct-irrelevant variance when targeting general intelligence. Such strategies for judging the "validity" or "interpretability" of IQs in measuring a unity concept exist (e.g., Flanagan & Kaufman, 2004; Groth-Marnat, 2003; Kaufman & Lichtenberger, 2006; Sattler, 2001). They are advocated, to some extent, in the interpretive manuals for every intelligence test battery included in our analyses, and they are also prominent in resources guiding practice. For example, in the text devoted to mental retardation in the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (APA, 2000), it is stated that

> when there is significant scatter in the subtest scores, the profile of relative strengths and weaknesses, rather than the mathematically derived full-scale IQ, will more accurately reflect the person's learning abilities. When there is a marked discrepancy across verbal and performance scores, averaging to obtain a full-scale IQ score can be misleading. (p. 42)

Despite these recommended practices, serious limitations to the most common methods of examining the influence of specific abilities on IQs (i.e., profile analysis) have been documented (Watkins, 2003; Watkins, Glutting, & Youngstrom, 2006).

Finally, in a broader context, if IQs for many individuals are not sufficiently exchangeable for all examinees, psychologists could develop policies and practices that rely on multiple measures of general intelligence (Meyer et al., 2001). When high-stakes decisions must be made, perhaps psychologists should choose two intelligence test batteries to administer to obtain IQs from different collections of tests drawing on different normative samples. We believe that this practice is not uncommon in school and clinic settings (Flynn, 2006; Reschly & Hosp, 2004), but we know of no recommendations for best practices in using multiple test batteries in this manner. On the other hand, rather than administer two intelligence test batteries, psychologists may choose to interpret and report multiple reliable and valid IQs yielded from different combinations of tests from the same battery. This practice would largely hold constant any effects stemming from differences in norming across test batteries. Regardless of how multiple IQs are obtained, to give their clients the benefit of doubt, psychologists should consider reporting the highest of these IQs (and its associated confidence interval) under most circumstances. However, more consideration of this practice is needed.

Conclusion

Well-trained, experienced, and insightful psychologists engaged in psychological assessments yielding IQs and measures of other important constructs in psychology should continue to rely on multiple methods of assessment and multiple sources of information to form their conclusions about their clients. Despite the apparent simplicity in the diagnostic criteria for conditions that require consideration of IQs, such as mental retardation, learning disorders, and intellectual giftedness, comprehensive assessments are needed to identify these conditions. In addition to the results yielded by intelligence test batteries, psychologists should attempt to integrate information about their clients' development and history; their clients' self-reports; and their clients' symptoms and impairment reported by knowledgeable others, as well as observations of their clients across settings when making diagnostic, eligibility, and intervention-related decisions for them.

References

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.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Belter, R. W., & Piotrowski, C. (2001). Current status of doctoral-level training in psychological testing. *Journal of Clinical Psychology, 57,* 717–726.

Bergeron, R., & Floyd, R. G. (2006). Broad cognitive abilities of children with mental retardation: An analysis of group and individual profiles. *American Journal of Mental Retardation, 111,* 417–432.

Bergeron, R., Floyd, R. G., & Shands, E. I. (in press). State eligibility guidelines for mental retardation: An update and consideration of part scores and unreliability of IQs. *Education and Training in Developmental Disabilities.*

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer.

Brody, N. (1997). Intelligence, schooling, and society. *American Psychologist, 52,* 1046–1050.

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 3,* 141–154.

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

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.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements.* New York: Wiley.

Elliott, C. D. (1990). *Differential Ability Scales.* San Antonio, TX: Psychological Corporation.

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

Floyd, R. G., Bergeron, R., McCormack, A. C., Anderson, J. L., & Hargrove-Owens, G. L. (2005). Are Cattell–Horn–Carroll (CHC) broad ability composite scores exchangeable across batteries? *School Psychology Review, 34,* 386–414.

Flynn, J. R. (1987). Massive gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171–191.

Flynn, J. R. (2006). Tethering the elephant: Capital cases, IQ, and the Flynn effect. *Psychology, Public Policy, and Law, 12,* 170–189.

Glutting, J. J., & Oakland, T. (1993). *Manual for the Guide to the Assessment of Test Session Behavior.* San Antonio, TX: Psychological Corporation.

Gottfredson, L. S. (2003). *g,* jobs and life. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 293–342). New York: Pergamon Press.

Groth-Marnat, G. (2003). *Handbook of psychological assessment* (4th ed.). Hoboken, NJ: Wiley.

Hammill, D. D., Brown, L., & Bryant, B. R. (1992). *A consumer's guide to tests in print* (2nd ed.). Austin, TX: PRO–ED.

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.

Individuals With Disabilities Education Act Amendments of 1997, Pub. L. No. 105–17, 20 U.S.C. 33, §§ 1400 *et seq.*

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Kamphaus, R. W., Petoskey, M. D., & Rowe, E. W. (2000). Current trends in psychological testing of children. *Professional Psychology: Research and Practice, 2,* 155–164.

Kanaya, T., Scullin, M. H., & Ceci, S. J. (2003). The Flynn effect and U.S. policies: The impact of rising IQ scores on American society via mental retardation diagnosis. *American Psychologist, 58,* 778–790.

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

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. (2004). *Kaufman Assessment Battery for Children, Second Edition, manual.* Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Lichtenberger, E. O. (2006). *Assessing adolescent and adult intelligence* (3rd. ed.). Needham Heights, MA: Allyn & Bacon.

Keith, T. Z., Fine, J. G., Taub, G. E., Reynolds, M. R., & Kranzler, J. H. (2006). Higher-order, multi-sample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children—Fourth Edition: What does it measure? *School Psychology Review, 35,* 108–127.

Keyes, D., Edwards, W., & Perske, R. (2002). People with mental retardation are dying, legally: At least 44 have been executed. *Mental Retardation, 40,* 235–242.

Koocher, G. P. (2003). IQ testing: A matter of life and death. *Ethics and Behavior, 13,* 1–2.

McConaughy, S. H., & Achenbach, T. M. (2004). *Manual for the Test Observation Form for Ages 2–18.* Burlington: University of Vermont, Research Center for Children, Youth, & Families.

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock–Johnson III* (Technical manual) Itasca, IL: Riverside.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128–165.

Murphy, L. L., Spies, R. A., & Plake, B. S. (2006). *Tests in print: VII.* Lincoln: University of Nebraska & Buros Institute of Mental Measurements.

National Research Council. (2002). *Mental retardation: Determining eligibility for Social Security benefits.* Washington, DC: National Academies Press.

Neisser, U. (Ed.). (1998). *The rising curve: Long-term gains in IQ and related measures.* Washington, DC: American Psychological Association.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw–Hill.

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.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20,* 33–65.

Reschly, D. J., & Hosp, J. L. (2004). State SLD identification policies and practices. *Learning Disability Quarterly, 27,* 197–213.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94,* 18–38.

Ryba, N. L., Cooper, V. G., & Zapf, P. A. (2003). Juvenile competence to stand trial evaluations: A survey of current practices and test usage among psychologists. *Professional Psychology: Research and Practice, 34,* 499–507.

Salvia, J., & Ysseldyke, J. (2003). *Assessment in special and remedial education* (8th ed.). Boston: Houghton Mifflin.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego, CA: Author.

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance, 15,* 187–211.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* New York: Macmillan.

Spitz, H. H. (1988). Wechsler subtest patterns of mentally retarded groups: Relationship to *g* and to estimates of heritability. *Intelligence, 12,* 279–297.

Sternberg, R. J. (Ed.). (1994). *Encyclopedia of human intelligence* (Vol. 2). New York: Macmillan.

Watkins, M. W. (2003). IQ subtest analysis: Clinical acumen or clinical illusion? *Scientific Review of Mental Heath Practice, 2,* 118–141.

Watkins, M. W., Glutting, J. J., & Youngstrom, E. A. (2006). Issues in subtest profile analysis. In D. P. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed., pp. 251–268). New York: Guilford Press.

Wechsler, D. (1991). *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). *Wechsler Intelligence Scale for Children—Fourth Edition.* San Antonio, TX: Psychological Corporation.

Wilson, M. S., & Reschly, D. J. (1996). Assessment in school psychology training and practice. *School Psychology Review, 25,* 9–23.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson Psychoeducational Battery—Third Edition.* Itasca, IL: Riverside.

Received June 27, 2006
Revision received April 25, 2007
Accepted May 21, 2007 ■