Correct Interpretation of Latent Versus Observed Abilities:

Implications From Structural Equation Modeling Applied to the WISC-III and WIAT Linking Sample

Hyeon-Joo Oh, *Educational Testing Service*
Joseph J. Glutting, *University of Delaware*
Marley W. Watkins, *The Pennsylvania State University*
Eric A. Youngstrom, *Case Western Reserve University*
Paul A. McDermott, *University of Pennsylvania*

In this study, the authors used structural equation modeling to investigate relationships between ability constructs from the *Wechsler Intelligence Scale for Children–Third Edition* (WISC-III; Wechsler, 1991) in explaining reading and mathematics achievement constructs on the *Wechsler Individual Achievement Test* (WIAT; Wechsler, 1992). Participants comprised the nationally stratified Linking sample (N = 1,116) of the WISC-III and WIAT. Relating the latent ability variables to the latent achievement variables showed that psychologists must go beyond g in order to meaningfully understand children's trait performance on the WISC-III. Results for reading indicated psychologists must pay attention to the constructs of g and Verbal Comprehension, and for mathematics, that they are obliged to consider g and Freedom From Distractibility. Results are discussed in terms of their theoretical, applied, and treatment-related implications.

Several lines of reasoning support the interpretation of IQ-test profiles. Each shares the premise that multidifferentiated constructions of intelligence provide greater insight into the nature and complexity of human ability and that by evaluating multiple abilities, psychologists gain greater diagnostic precision (Hale & Fiorello, 2002; Hale, Fiorello, Kavanagh, Hoeppner, & Gaither, 2001; Kaufman, 1994; Sattler, 2001). This perspective stands in direct opposition to a foundational rule of science: the law of parsimony, which holds that fewer variables are to be preferred whenever their explanatory power equals that of a more complex model. More formally, the law of parsimony states that "what can be explained by fewer principles is explained needlessly by more" (Occam's Razor; Jones, 1952, p. 620). Consequently, it is imperative for psychologists adopting the multidifferentiated perspective to demonstrate that their variables possess greater predictive or treatment validity than that obtainable from a more compact, or even a unitary, view of intelligence (Brody, 1985; Glutting, McDermott, Watkins, Kush, & Konold, 1997; Humphreys, 1962; Lubinski, 2000; McNemar, 1964; Messick, 1992).

The general intelligence (g) construct satisfies the law of parsimony. It is singular, and more important, the g-based score has excellent construct and criterion-related validity. An observed g-based score is also readily available on nearly all individually administered IQ tests. Examples include the Full Scale IQ (FSIQ) from the *Wechsler Intelligence Scale for Children–Third Edition* (WISC-III; Wechsler, 1991), the General Cognitive Ability (GCA) score from the *Differential Ability Scales* (DAS; C. D. Elliott, 1990), and the General IQ (GIQ) from the *Woodcock-Johnson III* (WJ-III; Woodcock, McGrew, & Mather, 2001).

The construct validity of g is well supported by factor analysis (Carroll, 1993; Gustafsson, 1989; Keith & Witta, 1997; Macmann & Barnett, 1994). More important, the greatest applied utility of the g-based score comes from its criterion-related validity. The utility of g-based scores, such as the FSIQ, GCA, and GIQ, in forecasting academic achievement is one of the most enduring findings in the fields of psychology and education (for reviews, see Board of Scientific Affairs of the American Psychological Association, 1996; Brody, 1985; Glutting, Adams, & Sheslow, 2000). Broadly speaking, g-based IQs correlate about .70 with standardized measures of achievement and .50 with grades in elementary school (Brody, 1985; Jensen, 1998). Because of range restrictions, ability–achievement correlations decrease as individuals advance through the educational system. Typical correlations between g and standardized high school achievement lie between .50 to .60; for college, coefficients vary between .40 and .50; and for graduate

Address: Hyeon-Joo Oh, Educational Testing Service, Rosedale Road, MS 15-L, Princeton, NJ 08541

school, correlations range between .20 and .40 (Brody, 1985; Jensen, 1998).

Large-scale studies also relate the importance of *g* in predicting less familiar criteria, such as aggression, delinquency, and crime (Caspi & Moffitt, 1993; Gordon, 1997; Wiegman, Kuttschreuter, & Baarda, 1992); health risks (Lubinski & Humphreys, 1997; Macklin et al., 1998); and income and poverty (Hunt, 1995; Murray, 1998). For instance, *g* covaries .20 to .60 with work performance, .30 to .40 with income, and approximately .30 with longevity (Brody, 1992, 1996; Gordon, 1997; Jensen, 1998; Lubinski, 2000). These correlates are especially interesting because they demonstrate how individual differences in *g* affect outcomes peripheral to education (Gottfredson, 1997; Lubinski, 2000).

At the same time, multiple systems have been advanced to interpret ability scores beyond *g*. Each assumes that discrete measures, such as subtest groupings or factor indexes, supply nontrivial information not contained in the *g*-based measure (cf. Kaufman, 1994; Sattler, 2001). Of these measures, factor scores are leading candidates for providing additional information. Factor scores are more valid than conceptual subtest groupings. Unlike inductively derived subtest organizations, such as Kaufman's (1994) and Sattler's (2001) groupings, factor scores retain considerable construct validity because they are formed *empirically* on the basis of factor analysis. Each factor score in a test battery (e.g., WISC-III, WJ-III) also accounts for more variance than that available from individual subtest scores. As a result, factor scores are more reliable than single subtest scores (as per the Spearman-Brown prophecy; Traub, 1991). Furthermore, because factor scores represent phenomena beyond the sum of subtest specificity, method variance, and measurement error, they potentially escape the myriad drawbacks that beset attempts to interpret subtest profiles (Glutting, McDermott, Konold, Snelbaker, & Watkins, 1999; McDermott, Fantuzzo, & Glutting, 1990; McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992; Watkins & Glutting, 2000; Watkins, Youngstrom, & Glutting, 2002). Consequently, factor scores promise the clinical benefits of ability differentiation while potentially avoiding problems plaguing the more common practice of subtest analysis.

Glutting, Youngstrom, Ward, Ward, and Hale (1997) examined the effectiveness of observed factor scores from the WISC-III, relative to the FSIQ, in predicting performance on the *Wechsler Individual Achievement Test* (WIAT; Wechsler, 1992). Data were examined via multiple regression analysis (MRA). Following longstanding methodologies for investigating the incremental validity of observed scores (cf. Humphreys, 1962; Lubinski & Dawis, 1992; McNemar, 1964; Messick, 1992), hierarchical MRA was employed where the parsimonious FSIQ was entered at the first step of the analysis, followed at the second step by the WISC-III's four factor scores (Verbal Comprehension, Perceptual Organization, Freedom From Distractibility, and Processing Speed). Dependent variables were Reading, Mathematics, Language, and Writing composites from the WIAT. Results, using both referred and nonreferred samples, showed that WISC-III factor scores failed to provide a substantial increase to the prediction of achievement after partialling the observed *g*-based estimate (i.e., the FSIQ). Similar findings were reported by Youngstrom, Kogos, and Glutting (1999), using observed factor scores and the *g*-based estimate (GCA) from the DAS.

In contrast, several authors showed that ability factors make important contributions to the understanding of achievement beyond *g* (Keith, 1999; McGrew, Keith, Flanagan, & Vanderwood, 1997). These authors argued that the Wechsler scales do not tap critical cognitive constructs and that MRA does not allow for a simultaneous analysis of general and specific abilities. To correct these perceived faults, they applied structural equation modeling (SEM) to the *Woodcock-Johnson Psycho-educational Battery–Revised* (WJ-R; Woodcock & Johnson, 1989). For example, McGrew et al. examined relationships among *g* and specific cognitive abilities with general and specific reading and mathematics skills. Results indicated that 40% of the variance in overall reading achievement was directly attributable to *g*. However, for younger children, 11% of the achievement in Letter–Word Identification was due to the specific ability of *Ga* (auditory processing) and 22% of the variation in Passage Comprehension was directly attributable to *Gc* (crystallized intelligence)—and these effects remained after *g* was partialled. Results therefore revealed that certain factor-based abilities were able to predict achievement above and beyond *g*.

Keith (1999) extended McGrew et al.'s (1997) study by investigating the effects of factor- and *g*-based abilities from the WJ-R for African American, Hispanic, and Caucasian students. Dependent variables were general and specific measures of reading and mathematics. Results were similar across ethnic groups: *g* accounted for a substantial proportion of variance in overall reading and mathematics achievement, but specific cognitive factors also contributed to the prediction of specific reading and mathematics outcomes. For example, up to 19% of the achievement variance in the Calculation criterion was due to the Processing Speed factor and around 11% of the variance in Letter–Word Identification criterion was due to the specific ability of *Ga*.

Disparities among outcomes from Glutting, Youngstrom, et al. (1997), Youngstrom et al. (1999), McGrew et al. (1997), and Keith (1999) may be a consequence of using different tests and different samples. However, it is more likely that the divergence is due to (a) the type of variables examined and (b) the statistical methodology employed. Glutting et al. and Youngstrom et al. examined observed factor scores, which are the standard scores (e.g., *M*s = 100, *SD*s = 15) psychologists interpret on ability and achievement tests. As a consequence of investigating observed scores, Glutting et al. and Youngstrom et al. employed MRA as their statistical methodology.

On the other hand, McGrew et al. (1997) and Keith (1999) based their conclusions on SEM. In SEM, interest is focused more on constructs than on observed scores. Thus, SEM methodology provides results that are best interpreted as relationships between underlying latent traits (i.e., constructs), while MRA concentrates on observed scores. Although related, latent traits and observed scores are not identical. Predictively, observed factors scores are more likely to result in problems as a consequence of multicolinearities (i.e., the observed factor scores are highly intercorrelated) and/or singularities (i.e., the FSIQ in the WISC-III is formed directly from a large number of subtests contributing to the factor scores). Therefore, SEM more accurately evaluates the true, or causal, effects of one *construct* on another.

Keith (1999) suggested that SEM should be applied to cognitive instruments other than the Woodcock-Johnson scales. Surprisingly, despite widespread use of Wechsler's tests, SEM has not been employed with the WISC-III and the WIAT. The research presented in this article used SEM to investigate the relative importance of general- versus specific-ability constructs from the WISC-III in predicting reading and mathematics achievement on the WIAT. Additionally, the current study expanded the SEM methodology used by McGrew et al. (1997) and Keith to include analyses of the effects of both general and specific cognitive abilities on both general and specific reading and mathematics achievement.

Method

Participants and Instruments

The SEM analyses employed standard scores from the linking sample of the WISC-III and the WIAT (Wechsler, 1992). The sample ($N = 1,116$) ranged in age from 6 years 0 months through 16 years 11 months and was nationally representative within ±2% of the 1990 U.S. Census on the variables of age, gender, race/ethnicity, region of country, and parent education level. Ability constructs were based on standard scores for 12 WISC-III subtests, including the mandatory five Verbal and five Performance subtests (Wechsler, 1991, p. 5). The supplementary subtests of Digit Span and Symbol Search were included because they undergird the WISC-III's factor indexes. The alternative Mazes subtest was not used. Both Digit Span and Symbol Search are regarded as primary components in most interpretation systems, whereas Mazes is traditionally excluded (e.g., see Kaufman, 1994; Sattler, 2001). The WIAT contains eight subtests that can be aggregated into four composites: Reading, Mathematics, Language, and Writing. Like McGrew et al.'s (1997) and Keith's (1999) studies, the current investigation concentrated on outcomes in reading and mathematics. Therefore, standard scores (*M*s = 100, *SD*s = 15) from subtests underlying the WIAT's Reading (Basic Reading and Reading Comprehension) and Mathematics (Numerical Operations and Mathematics Reasoning) composites were used as the observed achievement measures.

Models

SEM allows researchers to specify a priori, direct, and indirect relationships among variables in a model. Given the potential complexity of findings, results are typically portrayed through figures. In the current analyses, ability variables and traits were placed to the leftside of figures and achievement variables and traits were placed to the right. Observed variables were enclosed in rectangles (i.e., the measured WISC-III and WIAT scores). All observed scores were assumed to be affected by latent traits/constructs, which were enclosed in ellipses. The observed ability and achievement variables were also assumed to be affected by other influences, such as measurement error and unique subtest variances. These sources of variation were symbolized by small circles.

Factor scores, by their very nature, are indeterminant. An infinite number of factor scores are possible for an individual because rotational procedures in factor analysis are indefinite and open the possibility to a never-ending number of solutions. Nevertheless, four factors were previously supported for the WISC-III (Keith & Witta, 1997; Wechsler, 1991). These factors also appear on the profile sheet of WISC-III protocols, and they represent the factors examiners typically interpret. Consequently, the current study specified the same four first-order latent traits for the WISC-III:

1. Verbal Comprehension (VC), composed of observed scores from the Information, Similarities, Vocabulary, and Comprehension subtests;
2. Perceptual Organization (PO), developed from subtest scores on Picture Completion, Picture Arrangement, Block Design, and Object Assembly;
3. Freedom From Distractibility (FD), assembled from scores on the Arithmetic and Digit Span subtests; and
4. Processing Speed (PS), which was made up of scores from Coding and Symbol Search.

In certain models, the four first-order cognitive constructs were assumed to be caused by the second-order *g* trait. This portion of the model represents a hierarchical confirmatory factor analysis (CFA) of WISC-III abilities. Validity for this CFA organization was previously supported using ability subtests from the WISC-III standardization sample (Keith, 1997; Keith & Witta, 1997).

The right side of figures represented both observed and latent structures for the WIAT. For all reading analyses, two first-order latent traits were specified. These traits had direct correspondences to achievement subtests in the WIAT: Basic Reading and Reading Comprehension. The first-order reading traits were also assumed to be caused by a single second-order latent dimension, Reading, which paralleled the WIAT's Reading composite. Similarly, for all mathematics analyses, two first-order latent traits were identified according to subtest scores in the WIAT: Number Operation and Mathematics Reasoning. The two first-level mathematics factors were specified to be caused by a single second-order latent dimension labeled General Mathematics. Unreliability was estimated for the Reading and Mathematics subtests by setting errors and unique variances to the estimated reliability of the subtest (as listed in the WIAT manual) subtracted by 1 and then multiplied by the variance for the test (Bollen, 1998).

Nine models were developed between ability and reading achievement and another nine were developed between ability and mathematics achievement. The two sets of models were identical, except for their focus (i.e., reading vs. mathematics). Each model is identified in the following sections.

Model 1: *g* to General Achievement. This model best satisfies the law of parsimony. A single ability construct (*g*) was used to account for relationships. In essence, this model was found by Glutting et al. (1997) and Youngstrom et al. (1999) to work best for observed scores from the WISC-III and the DAS. One second-order latent ability trait (*g*) was used to estimate relationships to general reading or general mathematics achievement (i.e., a single, second-order latent-achievement trait was specified).

Model 2: Specific Abilities to General Achievement. The second model posited that multiple abilities alone provide greater precision in understanding general achievement. The *g* construct was not included. Instead, specific abilities (the first-order latent traits of VC, PO, FD, & PS) were used to estimate relationships to general reading or general mathematics achievement (a single, second-order latent-achievement trait).

Model 3: *g* to Specific Achievement. Model 3 was a variant of Model 1. The difference is that the single second-order latent-ability trait (*g*) was used to estimate relationships to specific, rather than general, reading or mathematics achievement. In the case of the reading analysis, two first-order latent achievement traits were specified (Basic Reading and Reading Comprehension). For the mathematics analysis, two other first-order latent traits were specified (Number Operation and Mathematics Reasoning).

Model 4: Specific Abilities to Specific Achievement. Model 4 was a variant of Model 2. The fourth model posited that multiple abilities alone provide greater precision in understanding specific achievement. The difference between Models 2 and 4 is that specific abilities were used in Model 4 to estimate relationships to specific, rather than general, reading or mathematics achievement. For the reading analysis, two first-order latent-achievement traits were specified (Basic Reading and Reading Comprehension). Likewise, for the mathematics analysis, two first-order latent traits were specified (Number Operation and Mathematics Reasoning).

Model 5: *g* and VC to Specific Achievement. McGrew et al. (1997) and Keith (1999) found that both *g* and certain specific abilities were necessary to understand achievement processes. Therefore, Model 5 used *g* and one specific ability construct (VC) to estimate relationships to specific reading or mathematics achievement.

Model 6: *g* and PO to Specific Achievement. Model 6 was a modification of Model 5. Like Model 5, Model 6 employed *g*. The difference is that PO (vs. VC) served as the specific ability construct used to estimate relationships to specific reading or mathematics achievement.

Model 7: *g* and FD to Specific Achievement. Model 7 also was a modification of Model 5. Here, FD served as the specific ability construct used to estimate relationships to specific reading or mathematics achievement.

Model 8: *g* and PS to Specific Achievement. Model 8 was another modification of Model 5, wherein PS served as the specific ability construct used to estimate relationships to specific reading or mathematics achievement.

Model 9: General Ability to General Achievement and Specific Abilities to Specific Achievements. Model 9 was a variant of the most parsimonious (i.e., best) model Keith (1999) and McGrew et al. (1997) obtained when they used SEM to examine ability and achievement constructs from the WJ-R. For reading, Model 9 provided paths from *g* to general reading achievement and from some of the specific WISC-III abilities to specific WIAT achievement constructs (i.e., VC to Basic Reading and Reading Comprehension). The path from *g* to the general reading construct (i.e., Reading) suggested that *g* affected general reading achievement. This path also affected the specific reading constructs. In other words, Model 9 specified that the effects of *g* on Basic Reading and Reading Comprehension would be found indirectly through the general achievement construct of Reading. The paths from specific abilities to specific achievements, in turn, tested whether these specific abilities affect specific achievements in addition to the effect of *g* on Reading. For mathematics, Model 9 was also specified according to both general and specific achievements.

In addition to the above, nested models based on Gustaffson and Balke's (1993) methods were attempted because they allow clear statements about the independent contribution of each latent construct. Unfortunately, those models did not converge and produced improper solutions.

Procedure

All models employed subtest standard scores and were evaluated through the Analysis of Moment Structures (AMOS; Arbuckle & Wothke, 1999) using maximum likelihood (ML) estimation. Inasmuch as ML was developed under the multivariate normality assumption, that assumption was checked by examining Mahalanobis distances, skewness, and kurtosis of the observed variables. Skewness and kurtosis were appropriate with no extreme values, but the Mahalanobis distance p value was less than .001 for 9 cases. Therefore, the nine outliers were deleted, leaving 1,107 cases for analysis.

Several measures of fit exist for evaluating the quality of SEM models, each was developed under somewhat different theoretical frameworks, and each focuses on different components (cf. Kaplan, 2000). Multiple measures were reported for the present study to highlight different aspects of fit in addition to the chi-square statistic: goodness of fit index (GFI), adjusted goodness of fit index (AGFI), Tucker-Lewis index (TLI), comparative fit index (CFI), and the root mean square error of approximation (RMSEA). The GFI is similar to a squared multiple correlation in that it provides the amount of variance/covariance that can be explained by the model under consideration (Kline, 1998; Tanaka, 1993). The AGFI, by contrast, is analogous to a squared multiple correlation corrected for model complexity. Thus, the AGFI is useful for comparing competing models. The TLI and CFI are conceptually different from one another, but both measure fit by comparing a given hypothesized model to a null model that assumes no relationship among the observed variables. The difference is that the TLI is less subject to sample size influences (Kranzler & Keith, 1999). These four measures range between 0 and 1.00, with larger values reflecting better fit. Traditionally, values of .90 or greater are interpreted as evidence of appropriate fit (Bentler & Bonett, 1980). However, more recent literature suggests that better fitting models produce values around .95 (Hu & Bentler, 1999).

The RMSEA takes into account the error of approximation in the population. This index tells how well a studied model fits the population covariance matrix—if it is available. RMSEA values of less than .05 indicate good fit, and values as high as .08 present reasonable errors of approximation in the population (Browne & Cudeck, 1993). The AMOS program used for the current study tests for the closeness of fit. That is, it tests the hypothesis that the RMSEA provides a good fit in the population. Jöreskog and Sörbom (1996) suggested that the p value for this test should be greater than .50.

Model comparison is another key consideration in SEM. Two such criteria are the Akaike Information Criterion (AIC; Akaike, 1974, 1987) and the Expected Cross Validation Index (ECVI), based on the work of Browne and Cudeck (1993). The AIC and ECVI are used to select one or more models from a set of plausible depictions and identify those likely to perform best with future samples of the same size drawn from the population in the same way. Small values of AIC and ECVI are associated with a better fit of the implied models (Jöreskog & Sörbom, 1993). Expected ranges for the AIC and ECVI are not possible because they are cross-validation indices, with smaller relative values indicating better fits.

Although multiple fit indices are reported, decisions concerning which model best fit the data were based primarily on the AIC and ECVI. Some, but not all, models in the study were nested. The AIC and ECVI offer a choice between competing models regardless of the nested status. The chi-square ($\Delta\chi^2$), GFI, AGFI, and RMSEA statistics were considered as supplemental indices for comparing models.

Results

Reading Achievement

Table 1 presents measures of fit for the nine reading achievement models. Models 6 and 7 resulted in improper solutions

TABLE 1. Comparison of Model Fit Measures of General and Specific Abilities on the Reading Accounting for Measurement Error

Model	χ^2	*df*	$\Delta\chi^2$	*p*	GFI	AGFI	TLI	CFI	RMSEA	AIC	ECVI
1	427.61	72			0.95	0.92	0.94	0.96	0.07***	493.61	0.45
2	1884.10	73	1456.49[a]	< .001	0.79	0.69	0.71	0.77	0.15***	1948.10	1.76
3	491.64	73			0.94	0.91	0.93	0.95	0.07***	555.64	0.50
4	1910.13	71	1418.49[b]	< .001	0.78	0.67	0.70	0.77	0.15***	1978.13	1.79
5	389.15	70			0.95	0.93	0.95	0.96	0.06***	459.15	0.42
6						Improper solution					
7						Improper solution					
8	421.75	70	32.60[c]	< .001	0.95	0.92	0.94	0.96	0.07	491.75	0.45
9	362.74	70	26.41[d]	< .001	0.96	0.93	0.95	0.96	0.06**	432.74	0.39

Note. GFI = Goodness of Fit Index; AGFI = Adjusted Goodness of Fit Index; TLI = Tucker-Lewis Index; CFI = Comparative Fit Index; RMSEA = Root mean square error of approximation; AIC = Akaike information criterion; ECVI = Expected Cross-Validation Index.
[a] $\Delta\chi^2$ of 1456.49 = χ^2 of Model 2 – χ^2 of Model 1. [b] $\Delta\chi^2$ of 1418.49 = χ^2 of Model 4 – χ^2 of Model 3. [c] χ^2 of 32.60 = χ^2 of Model 8 – χ^2 of Model 5. [d] $\Delta\chi^2$ of 26.41 = χ^2 of Model 5 – χ^2 of Model 9.
$^{**}p$ = .001. $^{***}p$ < .000.

because standardized parameter estimates exceeded 1.00 (see Note 1). Model 9 supplied the best solution and its overall fit was good. This model combined effects from *g*, as well as those from two factor-based abilities (VC and PO) in predicting both general and specific reading achievements. Consequently, relationships among aptitude and reading constructs were complex and more than the singular *g* construct was necessary to fully understand them. The equation with the second best fit was Model 5, which also employed both general- and specific-ability constructs (i.e., *g* to general reading achievement and VC to specific reading achievement). Nevertheless, the difference in chi-square values between Models 9 and 5 was 26.41, indicating that Model 9 was significantly better than Model 5; therefore, Model 9 provided the best relative fit (see Note 2). Results therefore showed that a combination of general and specific abilities was necessary to provide a reasonable explanation of reading achievement.

Standardized parameter estimates for Model 9 (direct, indirect, and total effects) are displayed in Figure 1. Although *g* was an important vehicle for understanding reading constructs, the specific abilities of VC and PO were also essential. For instance, the path from VC to General Reading shows a value of .30 and means that for each standard deviation increase in VC, performance on the General Reading construct increased .30 standard deviations. Path coefficients are standardized to *M* = 0.0 and *SD* = 1.0 and are interpreted in terms of standard deviation units. Effect size guidelines have been proposed for interpreting of standardized regression weights (Betas; Pedhazur, 1982), where values between .05 and .10 represent small effect sizes, values between .11 to .25 equal medium effect sizes, and values .26 or greater denote large effect sizes. Accordingly, findings revealed large effect sizes for VC across three reading constructs: General Reading (β = .30), Basic Reading (β = .26), and Reading Comprehension (β = .28). More important, the contribution of VC remained after the effect of *g* was already controlled.

Path coefficients as large as those for VC must be considered important and practically significant. For instance,

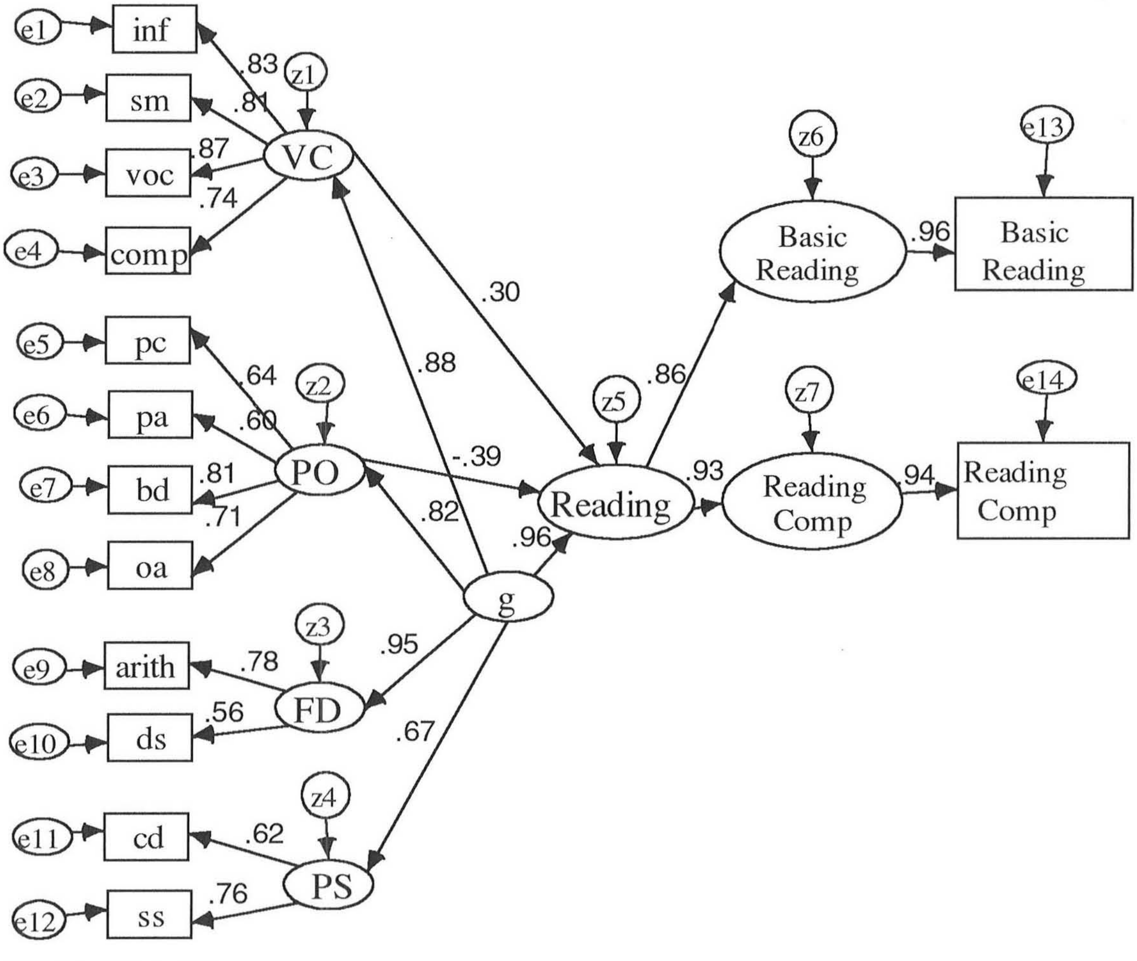

FIGURE 1. Effects of general and specific abilities on general reading achievement, counting random measurement error. *Note.* inf = information; sm = similarities; voc = vocabulary; comp = comprehension; pc = picture completion; pa = picture arrangement; bd = block design; oa = object assembly; VC = verbal comprehension; PO = perceptual organization; FD = freedom from distractibility; PS = processing speed.

outcomes for VC parallel findings from meta-analyses of such meaningful psychological relationships as the validity of screening tests in selecting job personnel (overall meta-analysis association = .27; Russell et al., 1994), the effect of Verbal scores from the Graduate Record Examination in predicting grade point averages (association = .28; Morrison & Morrison, 1995), and the overall effectiveness of psychotherapy (association = .32; Smith & Glass, 1977). Likewise, VC's contribution parallels such important medical findings as the effect of sleeping pills on reducing chronic insomnia (association = .30; Nowell et al., 1997); the relationship of stress tests in identifying heart disease (association = .30; Kwok, Kim, Grady, Segal, & Redberg, 1999), and the utility of mammograms in detecting breast cancer (association = .32; Mushlin, Kouides, & Shapiro, 1998).

Current VC outcomes rival, or exceed, effect sizes obtained during prior SEM studies of specific abilities (cf. Keith, 1999; McGrew et al., 1997). Therefore, in terms of relative utility, current findings clearly support inferences that VC made a meaningful contribution to the explanation of both general and specific reading constructs—and that it did so beyond levels offered by g.

Path coefficients in Figure 1 also make it clear that the FD and PS constructs provided no insight into children's reading achievement. Interestingly, results were contrary to expectations for the PO construct. Although PO had a moderate direct effect on General Reading ($\beta = -.39$) and moderate indirect effects on Basic Reading ($\beta = -.34$) and Reading Comprehension ($\beta = -.36$), all three effects were *negative*. The negative relationship is intriguing and raises a number of possibilities. The most likely appears to be the presence of multicolinearity. Thus, findings revealed that once the contribution of *g* and VC were controlled, for higher scores on PO were correlated with lower performance on General Reading, Basic Reading, and Reading Comprehension. The current findings hold important interpretive implications: Clinicians should pay little to no attention to children's performance on the ability constructs of PO, FD, and PS when trying to explain children's reading achievement. Instead, concentration should be confined to interpreting two WISC-III constructs—*g* and VC.

Table 2 offers further insights into the relative contributions made by *g*, VC, and PO. The most important rows are labeled "total" because they represent combined effects (i.e., total effect = direct effect + indirect effect). As previously established, unlike the negative contribution of PO, VC had an important positive effect on reading outcomes. At the same time, Table 2 makes it clear that g was at least three times more important than VC in explaining reading achievement. For example, *g* was 3.3 times more important than VC in understanding performance on the General Reading construct (total effect of *g* = .96; total effect of VC = .30: .96/.30 = 3.3); 3.0 times more important than VC in understanding Basic Reading outcomes (total effect of *g* = .78; total effect of VC = .26: .78/.26 = 3.0); and 3.0 times more important than VC in understanding Reading Comprehension (total effect of *g* = .84; total effect of VC = .28: .84/.28 = 3.0). Thus, the interpretive implication is clear: g is the most important construct in explaining reading performance, and clinicians need to give 3 times as much credence to the *g* factor as to interpretations from VC.

TABLE 2. Direct and Indirect Effects of WISC-III Traits in Predicting WIAT Reading Outcomes for the Best-Fitting Model

	WIAT outcome		
WISC-III predictor	General Reading	Basic Reading	Reading Comprehension
g			
Direct	0.96	0.00	0.00
Indirect	−0.05	0.78[a]	0.84
Total	0.91	0.78	0.84
Verbal Comprehension			
Direct	0.30	0.00	0.00
Indirect	0.00	0.26	0.28
Total	0.30	0.26	0.28
Perceptual Organization			
Direct	−0.39	0.00	0.00
Indirect	0.00	−0.34	−0.36
Total	−0.39	−0.34	−0.36

Note. WISC-III = *Wechsler Intelligence Scale for Children–Third Edition* (Wechsler, 1991); WIAT = *Wechsler Individual Achievement Test* (Wechsler, 1992).

[a] The indirect effect of *g* on Basic Reading is calculated using path coefficients provided in Figure 1 as follows: (coefficient from *g* to Verbal Comprehension × coefficient from Verbal Comprehension to Reading × coefficient from Reading to Basic Reading) + (coefficient from *g* to Perceptual Organization × coefficient from Perceptual Organization to Reading × coefficient from Reading to Basic Reading) + (coefficient from *g* to Reading × coefficient from Reading to Basic Reading). Thus, the resulting coefficient equals .78 (i.e., .88 × .30 × .86 + [.82 × −.39 × .86] + .96 × .86]).

Mathematics Achievement

Table 3 presents fit indices for the nine mathematics models. Like that for reading, standardized parameter estimates exceeded 1.00 for several models (5–7) and caused improper solutions (see Note 3). Results across analyses showed that Model 9 was best and that its level of fit was acceptable. This model combined effects from *g* and three factor-based abilities (VC, PO, and FD) in predicting both general and specific mathematics achievement. Therefore, results serve to emphasize the need to move beyond *g* when explaining mathematics achievement. Model 8 (*g* to general mathematics achievement and PS to specific mathematics achievements) provided the second-best fit after the variance of z3 was constrained to 0 (see Note 4). Table 3 shows that the difference in chi-aquare values between Model 9 and Model 8 was statistically significant but the effect size was small, $\Delta\chi^2(2) = 24.08, p < .001$.

Overall, results indicated that Model 9 offered the best fit. Standardized parameter estimates for this model (direct, indirect, and total effects) are displayed in Figure 2. General ability had a large direct effect on General Mathematics (β = .49) and an indirect effect on each specific mathematics construct; that is, *g* had an appreciable indirect effect on Number Operation (β = .79) and on Mathematics Reasoning (β = .88). With respect to the factor-based abilities, results revealed that VC and PS did not add to the explanation of general and specific mathematics constructs. PO evidenced moderate effect sizes on the mathematics constructs (highest β = –.26). However, against expectations, all of the effects were negative. Alternatively, FD showed a positive and appreciable direct effect on General Mathematics (β = .72) and positive and appreciable indirect effects on Number Operation (β = .63) and Mathematics Reasoning (β = .70), beyond *g*.

Table 4 further clarifies the relative contributions of *g*, VC, PO, and FD to the explanation of mathematics. Like that for reading, total effect sizes for *g* (i.e., its direct + indirect effects) were large for all three mathematics criteria: .92 to General Mathematics, .80 to Number Operation, and .90 to Mathematics Reasoning. By contrast, effect sizes for VC were negligible (–.05 to General Mathematics, –.04 to Number Operation, and –.04 to Mathematics Reasoning). Although effect sizes for PO were moderate, they also were negative and theoretically incongruent (–.26 to General Mathematics, –.23 to Number Operation, and –.26 to Mathematics Reasoning). Like that for the best reading model, the negative relationship between PO and general mathematics (i.e., Math) is intriguing and raises a number of possibilities. The most likely appears to be the presence of multicolinearity.

Only FD made a meaningful and theoretically congruent contribution to mathematics achievement above levels afforded by *g*. Effect sizes for FD were both theoretically congruent and large (.72 to General Mathematics, .63 to Number Operation, and .70 to Mathematics Reasoning). Consequently, outcomes reveal that clinicians need to pay little to no attention to children's performance on the ability constructs of VC, PO, and PS when trying to explain children's mathematics achievement. Instead, interpretations should be limited to *g* and FD.

Unlike reading outcomes, for which *g* was clearly superior to VC, effect sizes for mathematics showed that FD's contribution rivaled levels supplied by *g*. For example, *g* was only 1.27 times more important than FD in understanding performance on General Mathematics (total effect of *g* = .92; total effect of FD = .72: .92/.72 = 1.27); 1.25 times more important than FD in explaining Number Operation (total effect of *g* = .79; total effect of FD = .63: .79/.63 = 1.25); and 1.26

TABLE 3. Comparison of Model Fit Measures of General and Specific Abilities on the Mathematics Accounting for Measurement Error

Model	χ^2	*df*	$\Delta\chi^2$	*p*	GFI	AGFI	TLI	CFI	RMSEA	AIC	ECVI
1	385.58	73			0.95	0.93	0.95	0.96	0.06**	449.58	0.41
2	1871.50	73	1485.92[a]	< .001	0.73	0.69	0.72	0.76	0.15***	1935.50	1.75
3	440.55	73			0.95	0.92	0.94	0.95	0.07***	504.55	0.46
4	1843.21	71	1402.66[b]	< .001	0.79	0.68	0.72	0.78	0.15***	1911.21	1.73
5						Improper solution					
6						Improper solution					
7						Underidentification					
8	365.83	71	19.75[c]	< .001	0.96	0.93	0.95	0.96	0.06**	433.83	0.39
9	341.75	69	24.08[d]	< .001	0.96	0.94	0.96	0.97	0.06*	413.75	0.37

Note. GFI = Goodness of Fit Index; AGFI = Adjusted Goodness of Fit Index; TLI = Tucker-Lewis Index; CFI = Comparative Fit Index; RMSEA = Root mean square error of approximation; AIC = Akaike information criterion; ECVI = Expected Cross-Validation Index.
[a] $\Delta\chi^2$ of 1485.92 = χ^2 of Model 2 – χ^2 of Model 1. [b] $\Delta\chi^2$ of 1402.66 = χ^2 of Model 4 – χ^2 of Model 3. [c] $\Delta\chi^2$ of 19.75 = χ^2 of Model 1 – χ^2 of Model 8. [d] $\Delta\chi^2$ of 24.08 = χ^2 of Model 8 – χ^2 of Model 9.
**p* = .05. ***p* = .001. ****p* < .000.

times more important than FD in describing Mathematics Reasoning (total effect of *g* = .88; total effect of FD = .70: .88/.70 = 1.26). FD therefore made a large contribution to the explanation of mathematics achievement after the effect of *g* was controlled, and its explanatory power rivaled levels supplied by *g*. The interpretive implication is clear: Clinicians need to give equal weight to *g* and FD when explaining children's mathematics achievement. Given the importance of FD to mathematics achievement, it is disconcerting to note the absence of this construct in the recently released *Wechsler Intelligence Scale for Children–Fourth Edition* (WISC-IV; Wechsler, 2003).

Age Replications

McGrew et al. (1997) and Keith (1999) observed developmental trends during SEM analyses of the WJ-R. Replications were also attempted for four age groups from the WISC-III and WIAT Linking sample: 6 through 9 years, 10 through 13 years, and 14 through 17 years. Age-level replications failed to improve on the overall models for either reading or mathematics. Consequently, unlike that for the WJ-R, no meaningful developmental trends were observed for the WISC-III and the WIAT.

Discussion

Current results with the WISC-III and WIAT extend our knowledge about complex relationships between abilities and achievement. The results hold three sets of implications: theoretical, applied, and treatment-related.

Theoretical Implications

The current study based its findings on SEM, which is a multivariate technique designed to identify relationships among latent traits (i.e., constructs). Findings for both the reading and mathematics criteria make it clear that psychologists must go beyond *g* in order to meaningfully understand children's performance on the WISC-III. At the same time, results demonstrated psychologists should not give equal weight to all constructs in the WISC-III. For instance, when attempting to explain children's reading achievement on the WIAT, psychologists should limit interpretations to just two constructs: *g* and VC. No explanatory increase is obtained from PO, FD, or PS, and examining these traits in relationship to children's reading levels is simply a matter of overinterpretation. Similarly, when explaining children's mathematics achievement, psychologists should confine interpretations to just *g* and FD

TABLE 4. Direct and Indirect Effects of WISC-III Traits in Predicting WIAT Mathematics Outcomes for the Best-Fitting Model

	WIAT outcome		
WISC-III predictor	General Math	Number Operation	Math Reasoning
g			
Direct	0.49	0.00	0.00
Indirect	0.43	0.79[a]	0.88
Total	0.92	0.79	0.88
Verbal Comprehension			
Direct	−0.05	0.00	0.00
Indirect	0.00	−0.04	−0.05
Total	−0.05	−0.04	−0.05
Perceptual Organization			
Direct	−0.26	0.00	0.00
Indirect	0.00	−0.23	−0.25
Total	−0.26	−0.23	−0.25
Freedom From Distractibility			
Direct	0.72	0.00	0.00
Indirect	0.00	0.63	0.70
Total	0.72	0.63	0.70

Note. WISC-III = *Wechsler Intelligence Scale for Children–Third Edition* (Wechsler, 1991); WIAT = *Wechsler Individual Achievement Test* (Wechsler, 1992).

[a] The indirect effect of *g* on Number Operation is calculated using path coefficients provided in Figure 2 as follows: (coefficient from *g* to Verbal Comprehension × coefficient from Verbal Comprehension to Math × coefficient from Math to Number Operation) + (coefficient from *g* to Perceptual Organization × coefficient from Perceptual Organization to Math × coefficient from Math to Number Operation) + (coefficient from *g* to Freedom From Distractibility × coefficient from Freedom From Distractibility to Math × coefficient from Math to Number Operation) + (coefficient from *g* to Math × coefficient from Math to Number Operation). Thus, the resulting coefficient equals .79 (i.e., [.88 × −.05 × .87] + [.82 × −.26 × .87] + [.94 × .72 × .87] + [.49 × .87]).

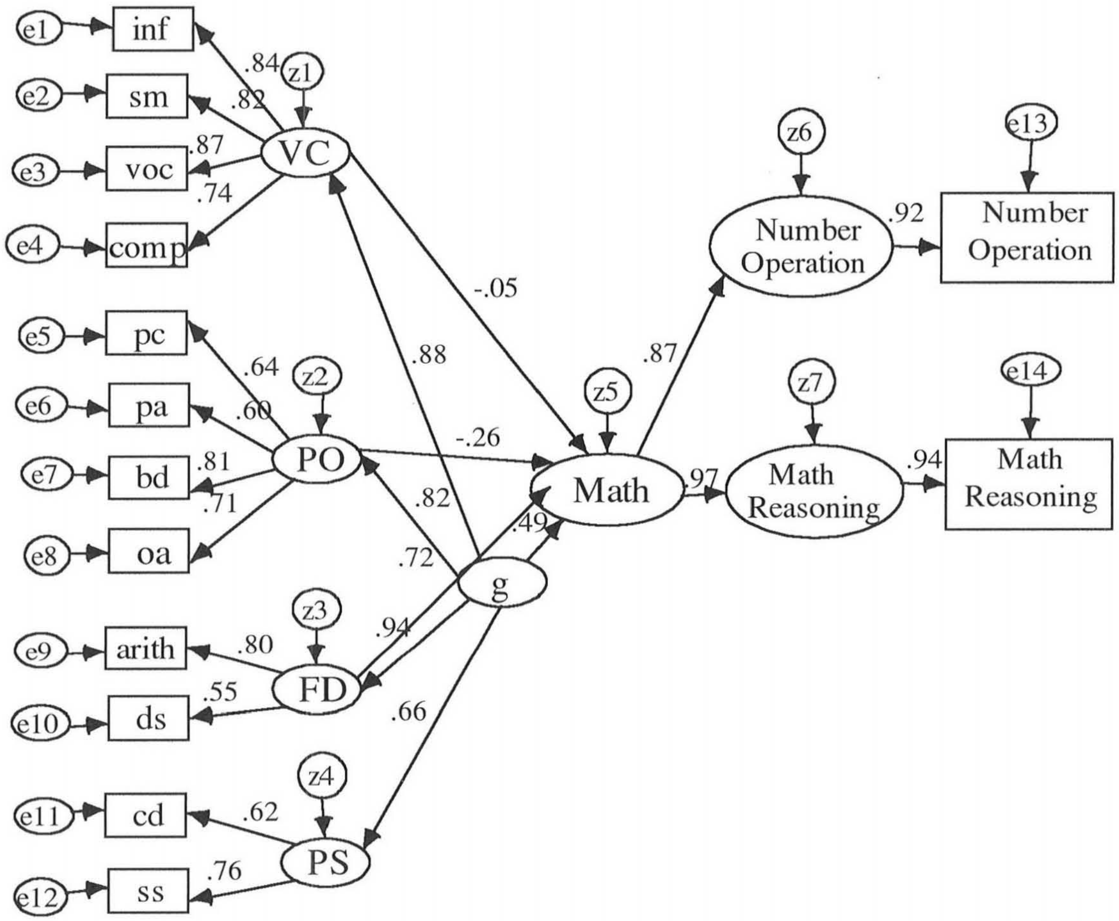

FIGURE 2. Effects of general and specific abilities on general mathematics achievement, counting random measurement error. *Note.* inf = information; sm = similarities; voc = vocabulary; comp = comprehension; pc = picture completion; pa = picture arrangement; bd = block design; oa = object assembly; vc = verbal comprehension; po = perceptual organization; FD = freedom from distractibility.

and ignore the VC, PO, and PS constructs. Thus, current outcomes strongly indicate that psychologists should look no further than the WISC-III constructs of *g*, VC, and FD when attempting to explain two of the most crucial outcomes in education: reading and mathematics achievement.

Current findings with the WISC-III and WIAT are consonant with outcomes obtained by Keith (1999) and McGrew et al. (1997) with the WJ-R. Similar conclusions were reached by Kuusinen and Leskinen (1988), as well as by Gustafsson and Balke (1993), with other measures of ability and achievement. When general and specific ability constructs are compared to general and specific achievement constructs, *g* usually accounts for the largest proportion of variance in achievement. However, additional achievement variance is explained by specific cognitive constructs. As noted by Detterman (2002), *g* only accounts for 25% to 50% of the variance in achievement outcomes, leaving 50% to 75% of the variance to be explained by other constructs. Likewise, "no one believes that *g* is the only construct needed to describe individual differences in intelligence" (Brody, 2002, p. 122). The findings obtained here in regard to the WISC-III need to be replicated and extended to the WISC-IV.

Applied Implications

Psychologists would be incorrect to assume that they can apply the current findings to their day-to-day assessments. For example, when examining for reading problems, psychologists might take these results to mean that they should limit interpretations to just the FSIQ and Verbal Comprehension Index of the WISC-III. However, even this restricted set of interpretations is probably too much.

To understand why overinterpretation is likely, psychologists must recognize that the observed scores obtained during routine clinical assessments are very different than the latent traits (i.e., constructs) derived by SEM. Observed scores are standard scores, such as the FSIQ, Index scores, and subtest scores in the WISC-III. Observed items and scales often contain variance from several sources and from several levels of generality (Gustafsson, 1994; Ullstadius, Gustafsson, & Carlstedt, 2002). The FSIQ, for example, is a mixture of *g*, specific cognitive skills, and systematic error (Colom, Abad, Garcia, & Juan-Espinosa, 2002).

SEM, on the other hand, provides results that are best interpreted as relationships among pure constructs measured without error. SEM is a good method for testing theory, but it is less satisfactory for direct, diagnostic applications. The observed scores employed by psychologists contain measurement error, whereas latent SEM traits do not (i.e., reliability coefficients = 1.00). Basing diagnostic decisions on theoretically pure constructs is very difficult in practice. Even approximating the construct scores derived from SEM requires complex, tedious calculations.

A case study will help clarify distinctions between observed scores and latent constructs. The appendix provides the steps necessary to convert observed scores from the WISC-III into constructs represented in Model 9, the model with the best fit for reading. It does so in the context of showing how SEM can be employed to develop IQ–achievement discrepancies. The purpose of the appendix is heuristic. It is not meant to endorse the application of IQ–achievement discrepancies. There are many legitimate reservations psychologists might have about using them to diagnose learning disabilities (Aaron, 1997; Fletcher et al., 1998; Fuchs, Fuchs, & Speece, 2002; Siegel, 1998; Vellutino, Scanlon, & Lyon, 2000). Instead, the appendix is useful for highlighting four dissimilarities between observed scores and latent traits. First, even a cursory review of equations reveals that SEM constructs are not equivalent to observed scores. Second, constructs rank children differently than observed scores, and, as the correlation between observed scores increases, so does the change. Thus, children's relative position on constructs (e.g., VC) can be radically different than their standing on corresponding observed scores (the Verbal Comprehension Index). Third, as the appendix makes clear, construct scores are not readily available to psychologists. Fourth, it is possible to estimate construct scores. However, until the equations appear in computer-interpretation programs, or unless psychologists are willing to engage in laborious calculations, they will have to rely on observed scores.

Perhaps the most important finding here is that psychologists cannot directly apply results from SEM. Observed scores must first be converted to construct scores before outcomes can be translated into practical, everyday uses. This situation holds not only for ability and achievement tests but for all SEM findings, regardless of whether analyses are directed to personality variables (e.g., parent, teacher, and self-reports of psychopathology), neuropsychological test scores, results from memory experiments, and the like.

One question remains: What should psychologists do if they do not want to calculate WISC-III construct scores and/or they prefer to interpret observed scores? We previously demonstrated that the FSIQ accounted for the lion's share of WIAT variance and that observed factor scores failed to substantially increase this prediction (Glutting, Youngstrom, et al., 1997). Therefore, psychologists who interpret observed scores should follow the guidelines provided in our earlier study (Oh, 2002) and heed the law of parsimony.

Treatment-Related Implications

A prominent finding from this study, as well as from nearly all similar studies conducted across the last half century, is that *g* is an excellent predictor of achievement. At the same time, it is becoming fashionable to laud the predictive aspects of *g* while simultaneously lamenting its lack of treatment validity. This position was taken by a number of scholars during a recent miniseries in the *School Psychology Review* (cf. Canter, 1997; S. N. Elliott & Fuchs, 1997; Esters, Ittenbach, & Han, 1997; Flanagan & Genshaft, 1997; Lopez, 1997; Reschly, 1997). It is true that as a target of intervention, *g* is noticeably resistant to change. A well-known example is Head Start, in which children have demonstrated initial intervention gains in *g*, followed by subsequent losses (for reviews, see Clarke & Clarke, 1989; Jensen, 1989; Spitz, 1986).

Alternatively, we believe *g* has much to offer interventions—not as a direct target, but as a consequence of creating *treatment expectancies*. For example, 10 to 25-year follow-up studies stressed the importance of *g*, as both a risk and a protective factor for children with attention-deficit/hyperactivity disorder (ADHD; Klein & Manuzza, 1991; Loney, Kramer, & Milich, 1981; Manuzza, Gittelman-Klein, Bessler, Malloy, & LaPadula, 1993; Weiss & Hechtman, 1993). IQ not only predicted academic performance in high school for individuals with ADHD (the most common expectancy) but also served as a risk and/or protective factor across multiple peripheral outcomes past high school. IQ was a significant indicator of whether children with ADHD had positive family relationships after high school, self-evaluations of their emotional adjustment as adults, and objective data dealing with adult psychiatric diagnoses, work performance, and socialization. Clearly, this information holds substantial treatment implications: Children with ADHD who have lower overall IQs will require more intensive interventions than those with higher IQs. Indeed, one need look no further for treatment implications than graduate training programs in school and clinical child psychology, where faculty take great pride in selecting the best and brightest as the targets of their interventions. Thus, to say *g* has no treatment validity is to miss the mark.

Conclusion

The history of psychology and education is littered with "advancements," whose benefits were later diminished, or refuted, when held up to empirical scrutiny. Throughout, IQ testing has endured. It has withstood blistering attacks from critics, as well as improper use by advocates. If psychologists and other assessment specialists intend to accurately evaluate the abilities of children and adolescents, they must shift focus from popular practices to interpretations based on sound research. This means psychologists must begin to recognize fundamental distinctions between factor-based versus inductively derived subtest groupings, between observed scores versus ipsatized scores, and, as found here, between observed scores versus latent constructs. IQ tests are useful, but only if we interpret their scores correctly.

NOTES

1. The FD and PO factors in Models 6 and 7, respectively, resulted in improper solutions as a consequence of multicolinearities. Attempts were made to constrain one or more parameters in the two models in order to arrive at proper solutions (i.e., a parameter was constrained to 0). However, removing the FD and PO factors would be inappropriate because it is a viable construct in the WISC-III and their observed scores are frequently interpreted by psychologists. Therefore, the FD and PO constructs were not removed, and their presence in Models 6 and 7 resulted in improper solutions.
2. The degree of freedom of Models 5 and 9 is the same (i.e., *df* = 70). Therefore, it is not possible to present the *p* value.
3. The VC, FD, and PO factors in Models 5, 6, and 7, respectively, resulted in Heywood cases as a consequence of multicolinearities. Attempts were made to constrain one or more parameters in the two models in order to arrive at proper solutions (i.e., a parameter was constrained to 0). However, removing the VC, FD, and PO factors would be inappropriate because they are viable constructs in the WISC-III and their observed scores are frequently interpreted by psychologists. Therefore, the VC, FD, and PO constructs were not removed, and their presence in Models 5, 6, and 7 resulted in improper solutions.
4. An initial analysis of Model 8 found that the error variance of the Mathematics Reasoning trait was negative (Heywood case) due to multicolinearities among FD, PO, and the mathematics achievement subtests. Therefore, this variance was constrained to 0 in order to obtain proper solutions.

REFERENCES

Aaron, P. G. (1997). The impending demise of the discrepancy formula. *Review of Educational Research, 67,* 461–502.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–123.

Akaike, H. (1987). Factor analysis and AIC. Psychometrika, 52, 317–332.

Arbuckle, J. L., & Wothke, W. (1999). *AMOS 4.0: User's guide.* Chicago: Small Waters.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Board of Scientific Affairs of the American Psychological Association. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77–101.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Brody, N. (1985). The validity of tests of intelligence. In B. Wolman (Ed.), *Handbook of intelligence* (pp. 353–389). New York: Wiley.

Brody, N. (1992). *Intelligence.* San Diego, CA: Academic Press.

Brody, N. (1996). Intelligence and public policy. *Psychological Public Policy Law, 3/4,* 473–485.

Brody, N. (2002). *g* and the one–many problem: Is one enough? In *The nature of intelligence* (Novartis Foundation Symposium 233; pp. 122–135). New York: Wiley.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural models* (pp.136–162). Newbury Park, CA: Sage.

Canter, A. S. (1997). The future of intelligence testing in the schools. *School Psychology Review, 26,* 255–261.

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

Caspi, A., & Moffitt, T. E. (1993). When do individual differences matter? A paradoxical theory of personality coherence. *Psychological Inquiry, 4,* 247–321.

Clarke, A. M., & Clarke, A. D. B. (1989). The later cognitive effects of early intervention. *Intelligence, 13,* 289–297.

Colom, R., Abad, F. J., Garcia, L. F., & Juan-Espinosa, M. (2002). Education, Wechsler's full scale IQ, and g. *Intelligence, 30,* 449–462.

Elliott, C. D. (1990). *Differential ability scales.* San Antonio, TX: Psychological Corp.

Elliott, S. N., & Fuchs, L. S. (1997). The utility of curriculum-based measurement and performance assessment alternatives to traditional intelligence and achievement tests. *School Psychology Review, 26,* 224–233.

Esters, I. G., Ittenbach, R. F., & Han, K. (1997). Today's IQ tests: Are they really better than their historical predecessors? *School Psychology Review, 26,* 211–224.

Flanagan, D. P., & Genshaft, J. L. (1997). Issues in the use and interpretation of intelligence tests in the schools: Guest editors' comments. *School Psychology Review, 26,* 146–149.

Fletcher, J. M., Francis, D. J., Shaywitz, S. E., Lyon, G. R., Foorman, B. R., Stuebing, K. K., & Shaywitz, B. A. (1998). Intelligent testing and the discrepancy model for children with learning disabilities. *Learning Disabilities Research & Practice, 13,* 186–203.

Fuchs, L. S., Fuchs, D., & Speece, D. L. (2002). Treatment validity as a unifying construct for identifying learning disabilities. *Learning Disability Quarterly, 25,* 33–45.

Glutting, J. J., McDermott, P. A., Konold, T. R., Snelbaker, A. J., & Watkins, M. W. (1999). More ups and downs of subtest analysis: Criterion validity of the DAS with an unselected cohort. *School Psychology Review, 27,* 597–610.

Glutting, J. J., McDermott, P. A., Prifitera, A., & McGrath, E. A. (1994). Core profile types of the WISC-III and WIAT: Their development and application in identifying multivariate IQ-achievement discrepancies. *School Psychology Review, 23,* 619–639.

Glutting, J. J., McDermott, P. A., Watkins, M. W., Kush, J. C., & Konold, T. R. (1997). The base rate problem and its consequences for interpreting children's ability profiles. *School Psychology Review, 26,* 176–188.

Glutting, J. J., Youngstrom, E. A., Ward, T., Ward, S., & Hale, R. (1997). Incremental efficacy of WISC-III factor scores in predicting achievement: What do they tell us? *Psychological Assessment, 9,* 295–301.

Gordon, R. A. (1997). Everyday life as intelligence test. *Intelligence, 24,* 303–320.

Gottfredson, L. S. (1997). Intelligence and social policy. *Intelligence, 24,* 288–320.

Gustafsson, J.-E. (1989). Broad and narrow abilities in research on learning and instruction. In R. Confer, P. L. Ackerman, & R. Cudeck (Eds.), *Abilities, motivation, and methodology: The Minnesota symposium on learning and individual differences* (pp. 203–237). Hillsdale, NJ: Erlbaum.

Gustafsson, J.-E. (1994). Hierarchical models of intelligence and educational achievement. In A. Demetrious & A. Efklides (Eds.), *Intelligence, mind, and reasoning: Structure and development* (pp. 45–73). New York: Elsevier Science.

Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28,* 407–434.

Hale, J. B., & Fiorello, C. (2002). Beyond the academic rhetoric of *g*: Intelligence testing guidelines for practitioners, Part II. *Communiqué: Newspaper of the National Association of School Psychologists, 31*(3), 12–15.

Hale, J. B., Fiorello, C. A., Kavanagh, J. A., Hoeppner, J. B., & Gaither, R. A. (2001). WISC-III predictors of academic achievement for children with learning disabilities: Are global and factor scores comparable? *School Psychology Quarterly, 16*(1), 31–55.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist, 17*, 475–483.

Hunt, E. (1995). *Will we be smart enough?* New York: Russell Sage Foundation.

Jensen, A. R. (1989). Raising IQ without raising g? A review of "The Milwaukee Project: Preventing mental retardation in children at risk." *Developmental Review, 9*, 234–258.

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

Jones, W. T. (1952). *A history of western philosophy.* New York: Harcourt, Brace.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: User's reference guide.* Chicago: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1996). *Structural equation modeling.* Workshop presented for the NORC Social Sciences Research Professional Development Training Sessions, Chicago.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions. Advanced quantitative techniques in the social sciences series 10.* Thousand Oaks, CA: Sage.

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

Keith, T. Z. (1997). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 373–402). New York: Guilford Press.

Keith, T. Z. (1999). Effects of general and specific abilities on student achievement: Similarities and differences across ethnic groups. *School Psychology Quarterly, 14*, 239–262.

Keith, T. Z., & Witta, E. L. (1997). Hierarchical and cross-age confirmatory factor analysis of the WISC-III: What does it measure? *School Psychology Quarterly, 12*, 89–107.

Klein, R. G., & Mannuzza, S. (1991). Long-term outcome of hyperactive children. *Journal of the American Academy of Child and Adolescent Psychiatry, 30*, 383–387.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York: Guilford Press.

Kranzler, J. H., & Keith, T. Z. (1999). Independent confirmatory factor analysis of the Cognitive Assessment System (CAS): What does the CAS measure? *School Psychology Review, 28*, 117–144.

Kuusinen, J., & Leskinen, E. (1988). Latent structure analysis of longitudinal data on relations between intellectual abilities and school achievements. *Multivariate Behavioral Research, 23*, 103–118.

Kwok, Y., Kim, C., Grady, D., Segal, M., & Redberg, R. (1999). Meta-analysis of exercise testing to detect coronary disease in women. *American Journal of Cardiology, 83*, 660–666.

Loney, J., Kramer, J., & Milich, R. S. (1981). The hyperactive child grows up: Predictors of symptoms, delinquency and achievement at follow-up. In K. D. Gadow & J. Loney (Eds.), *Psychosocial aspects of drug treatment of hyperactivity* (pp. 381–415). Boulder, CO: Westview.

Lopez, R. (1997). The practical impact of current research and issues in intelligence interpretation and use of multicultural populations. *School Psychology Review, 26*, 249–254.

Lubinski, D. (2000). Scientific and social significance of assessing individual differences: "Sinking shafts at a few critical points." *Annual Review of Psychology, 51*, 405–444.

Lubinski, D., & Dawis, R. V. (1992). Aptitudes, skills, and proficiencies. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of I/O psychology* (Vol. 3; pp. 1–59). Palo Alto, CA: Consulting Psychological Press.

Lubinksi, D., & Humphreys, L. G. (1997). Incorporating general intelligence into epidemiology and the social sciences. *Intelligence, 24*, 159–201.

Macklin, M. L., Metzger, L. J., Litz, B. T., McNally, R. J., Lasko, N. B., Orr, S. P., & Pitman, R. K. (1998). Lower precombatant intelligence is a risk factor for posttraumatic stress disorder. *Journal of Consulting and Clinical Psychology, 66*, 323–326.

Macmann, G. M., & Barnett, D. W. (1994). Structural analysis of correlated factors: Lessons from the verbal-performance dichotomy of the Wechsler scales. *School Psychology Quarterly, 9*, 161–197.

Manuzza, S., Gittelman-Klein, R., Bessler, A., Malloy, P., & LaPadula, M. (1993). Adult outcomes of hyperactive boys: Educational achievement, occupational rank, and psychiatric status. *Archives of General Psychiatry, 50*, 565–576.

McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique of Wechsler theory and practice. *Journal of Psychoeducational Assessment, 8*, 290–302.

McDermott, P. A., Fantuzzo, J. W. Glutting, J. J., Watkins, M. W., & Baggaley, A. R. (1992). Illusions of meaning in the ipsative assessment of children's abilities. *The Journal of Special Education, 25*, 504–526.

McGrew, K. S., Keith, T. Z., Flanagan, D. P., & Vanderwood, M. (1997). Beyond *g*: The impact of Gf-Gc specific cognitive ability research on the future use and interpretation of intelligence tests in the schools. *School Psychology Review, 26*, 189–201.

McNemar, Q. (1964). Lost: Our intelligence? Why? *American Psychologist, 19*, 871–882.

Messick, S. (1992). Multiple intelligences or multilevel intelligence? *Psychological Inquiry, 3*, 365–384.

Morrison, T., & Morrison, M. (1995). A meta-analytic assessment of the predictive validity of the quantitative and verbal components of the Graduate Record Examination with graduate grade point average representing the criterion of success. *Educational and Psychological Measurement, 55*, 309–316.

Murray, C. (1998). *Income, inequality, and IQ.* Washington, DC: American Enterprise Institute.

Mushlin, A. I., Kouides, R. W., & Shapiro, D. E. (1998). Estimating the accuracy of screening mammography: A meta-analysis. *American Journal of Preventive Medicine, 14*, 143–153.

Nowell, P. D., Mazumdar, S., Buysse, D. J., Dew, M. A., Reynolds, C. F., III, & Kupfer, D. F. (1997). Benzodiazepines and zolpidem for chronic insomnia: A meta-analysis of treatment efficacy. *Journal of the American Medical Association, 278*, 2170–2177.

Oh, H. J. (2002). *Relative importance of general and specific abilities from the WISC-III in predicting achievement using SEM methodology.* Unpublished doctoral dissertation, University of Delaware, Newark.

Pedhazur, E. (1982). *Multiple regression in behavioral research* (2nd ed.). New York: Holt, Rinehart & Winston.

Reschly, D. J. (1997). Utility of individual ability measures and public policy choices for the 21st century. *School Psychology Review, 26*, 234– 241.

Russell, C. J., Settoon, R. P., McGrath, R. N., Blanton, A. E., Kidwell, R. E., Lohrke, F. T., et al. (1994). Investigator characteristics as moderators of personnel selection research: A meta-analysis. *Journal of Applied Psychology, 79*, 163–170.

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

Siegel, L. S. (1998). The discrepancy formula: Its use and abuse. In B. K. Shapiro, P. J. Accardo, & A. J. Capute (Eds.), *Specific reading disability: A view of the spectrum* (pp. 123–135). Timonium, MD: York Press.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*, 752–760.

Spitz, H. H. (1986). *The raising of intelligence: A selected history of attempts to raise retarded intelligence.* Hillsdale, NJ: Erlbaum.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. S. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp.10–39). Newbury Park, CA: Sage.

Thorndike, R. L. (1963). *The concepts of over- and underachievement.* New York: Teachers College Press.

Thorndike, R. L. (1994). G (Editorial). *Intelligence, 19,* 145–155.

Traub, R. E. (1991). *Reliability for the social sciences.* Newbury Park, CA: Sage.

Ullstadius, E., Gustafsson, J. E., & Corlstedt, B. (2002). Influence of general and crystallized intelligence on vocabulary test performance. *European Journal of Psychological Assessment, 18,* 78–84.

Vellutino, F. R., Scanlon, D. M., & Lyon, G. R. (2000). Differentiating between difficult-to-remediate and readily remediated poor readers: More evidence against the IQ–achievement discrepancy definition for reading disability. *Journal of Learning Disabilities, 33,* 223–238.

Watkins, M. W., & Glutting, J. J. (2000). Incremental validity of WISC-III profile elevation, scatter, and shape information for predicting reading and math achievement. *Psychological Assessment, 12,* 402–408.

Watkins, M. W., Youngstrom, E. A., & Glutting, J. J. (2002). Cautions concerning cross-battery assessment. *Communiqué of the National Association of School Psychologists, 30,* 8–16.

Wechsler, D. (1991). *Wechsler intelligence scale for children–Third edition.* San Antonio, TX: Psychological Corp.

Wechsler, D. (1992). *Wechsler individual achievement test: Manual.* San Antonio, TX: Psychological Corp.

Wechsler, D. (2003). *Wechsler intelligence scale for children–Fourth edition.* San Antonio, TX: Psychological Corp.

Weiss, G., & Hechtman, L. (1993). *Hyperactive children grown up* (2nd ed.). New York: Guilford Press.

Wiegman, O., Kuttschreuter, M., & Baarda, B. (1992). A longitudinal study of the effects of television viewing on aggressive and prosocial behavior. *British Journal of Social Psychology, 31,* 147–164.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III.* Itasca, IL: Riverside.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson tests of cognitive ability–Revised.* Chicago: Riverside.

Youngstrom, E. A., Kogos, J. L., & Glutting, J. J. (1999). Incremental efficacy of Differential Ability Scales factor scores in predicting individual achievement criteria. *School Psychology Quarterly, 14*(1), 26–39.

Appendix

Calculating IQ–Achievement Discrepancies From Latent Traits

Psychologists are familiar with the use of observed scores used during regression-discrepancy analysis. By contrast, if they want to employed constructs from the current study to predict reading achievement, they need to complete the following steps.

1. Convert all observed WISC-III subtest standard scores to *z* scores. The conversion is necessary because standard scores for the WISC-III's observed factors are expressed with means of 100 and standard deviations of 15. By contrast, the scales of constructs (i.e., VC, PO, FD, PS, and *g*) used in SEM are expressed as *z* scores (*M*s = 0.0, *SD*s = 1.0).

2. Estimate the covariance matrix for latent endogenous variables (i.e., VC, PO, FD, and PS).

$$\text{Covariance matrix of } \eta,\ \hat{\Sigma}(\eta) = \Gamma * \Phi * \Gamma' + \Psi$$

where η is an $m \times 1$ vector of endogenous latent variables, m is a number of endogenous variables, Γ is an $m \times k$ matrix of regression coefficients relating endogenous variables to exogenous variables, Φ is a covariance matrix of ξ, which is 1, because there are only exogenous variables in the current analysis, ξ is latent exogenous variable, Γ' is the transpose of Γ, Ψ is an $m \times m$ covariance matrix of, and ζ, is an $m \times 1$ vector of disturbance terms.

3. Estimate first-order factor scores, which are latent endogenous variables (i.e., η).

$$\eta = \hat{\Sigma}(\eta) * \Lambda'_y * \Sigma^{-1} * y$$

where Λ'_y is the transpose of $m \times p$ matrix of factor loadings, p is a number of indicators of η, y is the observed indicators of η, and Σ^{-1} is the inverse of population covariance matrix.

4. Estimate the second-order factor score, which is a latent exogenous variable (i.e., ξ).

$$\xi = \Phi * \Gamma' * \hat{\Sigma}(\eta)^{-1} * \eta$$

5. Convert the estimated ηs and ξ back to a scale score of mean of 100 and standard deviation of 15.

6. Estimate expected reading score using the familiar univariate regression equation (see Thorndike, 1963, for the equation and Glutting, McDermott, Prifitera, & McGrath, 1994, for an applied discussion). For example, for reading achievement, the best-fitting model was Model 9. Thus, the output of Model 9 can be used with the following equation to predict reading achievement on the Basic Reading and Reading Comprehension traits from the WIAT:

$$\text{EXP. ACH}_{\text{Reading}} = (\beta_{\text{RVC}} * \eta_{\text{VC}}) + (\beta_{\text{RPO}} * \eta_{\text{PO}}) + (\gamma_{\text{R}g} * \xi_g) + \zeta_{\text{R}}$$

where $\gamma_{\text{R}g}$ is the regression coefficient of *g* on Reading, ξ_g is the estimated exogenous variable of *g*, ζ_{R} is the disturbance term of Reading, β_{RVC} is the regression coefficient of Verbal Comprehension on Reading, η_{VC} is the estimated endogenous variable of Verbal Comprehension, β_{RPO} is the regression coefficient of PO on Reading, η_{PO} is the estimated endogenous variable of PO. Equations used to estimate the covariance matrix are as follows: (a) endogenous variables, (b) first-order factors, and (c) second-order factor scores, which were obtained from Bollen (1989) and Dr. Edward Rigdon (personal communication, January 3, 2002).

Case Example

Assume a child, John, obtained a standard score of 75 on the WIAT Reading Composite. His obtained standard scores were all 10 on subtests comprising the observed Perceptual Organization, Freedom From Distractibility, and Processing Speed factors of the WISC-III. By contrast, his observed standard score on the Information subtest was 7; his standard score on Vocabulary was 7, his scores on Similarities was 7; and his score on Comprehension was 9. Therefore, his *z* scores were as follows: −1.0 for Information, −1.0 for Vocabulary, −1.0 for Similarities, −0.33 for Comprehension, and 0 for the other subtests (Step 1). Converting produces the elements of vector *y*, which are the observed indicators of h from the equation (Step 2).

They are as follows:

$$\underset{12\times 1}{y} = \begin{bmatrix} -1 \\ -1 \\ -1 \\ -.33 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Estimates of the first- and second-factor scores using the Excel application are shown below (Steps 2–4):

$$\underset{4\times 1}{\eta} = \begin{bmatrix} -0.70417 \\ -0.24744 \\ -0.35017 \\ -0.25947 \end{bmatrix}$$

$$\xi = -0.42614$$

Converted estimates of ηs and ξ to means of 100 and standard deviations of 15 produces the value of 89.43 (i.e., 89.43 = −0.70417 × 15 + 100) for Verbal Comprehension, 96.29 for Perceptual Organization, 94.75 for Freedom From Distractibility, 96.11 for Processing Speed, and 96.61 for the Full Scale IQ (Step 5).

John's expected reading achievement score is 95.01 (Step 6). Estimates of β_{RVC}, β_{RPO}, γ_{Rg}, and, ζ_{R} (.30, −.39, .96, and 12.99, respectively) were obtained from the AMOS output and take into account random measurement error. Inserting these values into the predicted-achievement equation results in the following:

$$\text{EXP. ACH}_{\text{Reading}} = (.30 * 89.43) + (-.39 * 96.29) + (.96 * 96.61) + 12.99$$
$$= 95.01$$

John's obtained composite Reading Index on the WIAT was 75. This score is below his expected reading score as estimated from his obtained, multiple-ability trait scores. The resultant *z* score for the discrepancy between obtained and expected achievement is 3.91 (see the following equation).

$$z = \frac{|\text{EXP. ACH} - \text{OBT. ACH}|}{SD_{\text{ACH}} \sqrt{1 - r^2_{\text{IQ/ACH}}}}$$

This value corresponds to a prevalence of less than 0.01% and suggests that John may be eligible for classification as learning disabled.