Article

What If We Took Our Models Seriously? Estimating Latent Scores in Individuals

Journal of Psychoeducational Assessment
31(2) 186–201

DOI: 10.1177/0734282913478046
jpa.sagepub.com
SAGE

W. Joel Schneider[1]

Abstract

Researchers often argue that the structural models of the constructs they study are relevant to clinicians. Unfortunately, few clinicians are able to translate the mathematically precise relationships between latent constructs and observed scores into information that can be usefully applied to individuals. Typically this means that when a new structural model supplants a rival model, clinicians have only vague (and often incorrect) notions about how score interpretations should change. Fortunately, it is possible to estimate latent scores from observed scores in a rigorous manner. More important, paying attention to the confidence intervals around those estimates can assist clinicians' intuitions about what can (and cannot) be known with precision about a person's abilities. These methods are illustrated with structural models of the WISC-IV (Wechsler Intelligence Scale for Children—Fourth Edition). A free user-friendly spreadsheet that automates these procedures is available from the author.

Keywords

cognitive assessment, confirmatory factor analysis, estimated latent scores

It is almost certain that the final models in the target articles are not exactly final. If some psychology-minded philanthropist were to commission a series of lavishly funded studies that systematically linked all of the cognitive and academic tests fit for clinical use, better models would certainly emerge. However, the models presented in the target articles (Weiss, Keith, Zhu, & Chen, 2013a, 2013b) are reasonably close to the limits of how far single-battery confirmatory factor analyses (CFAs) of this sort can take us in understanding which abilities are measured by the Wechsler Intelligence Scale for Children—Fourth Edition (WISC-IV) and the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV). I might wish to tweak the models here or there: I would have liked to have seen a comparison of the final model with an analogous nested factor model. I am concerned about the inability to separate Gf and *g*. I would prefer to have left the small but significant cross-loadings in the models. I would also like to have seen the effects of separating Digits Forward and Digits Backward. Nevertheless, I am prepared to take a leap of faith and declare the models to be excellent and ready to be used with individuals in applied settings.

[1]Illinois State University, Normal, IL, USA

Corresponding author:
W. Joel Schneider, Department of Psychology, Illinois State University,
Campus Box 4620, Normal, IL 61790-4620, USA.
Email: wjschne@ilstu.edu

What I mean by this is not what is typically meant by such a statement. Usually it means eyeballing cognitive profiles and intuiting whether they match profiles imagined to be typical of a model. No, I mean something much more precise. When we have good models, we should use them and, within reason, get as much mileage from them as we can. For clinicians who agree with me that these models are reasonably close to the best available models for the WISC-IV and WAIS-IV, I will present a method of using these models to interpret individual profiles. The method is mathematically rigorous, but it is not mechanical. That is, there is room for the clinician to think but flights of fancy are constrained by confidence intervals.

A Bit of History and Some Comments About *g*

Spearman's (1904, 1927) two-factor theory is an attempt to explain individual differences in cognitive abilities with only two types of abilities: general and specific. For several decades, Spearman believed that there was only a single general ability and many test-specific abilities. Until evidence to the contrary became undeniable, Spearman did not believe that there were "group factors." Group factors are abilities that have substantial influences on some, but not all tests. Examples of group factors are inductive reasoning, visual-spatial ability, short-term memory, and processing speed.

There is a reason that Spearman's two-factor theory survived as long as it did. In most factor analyses of cognitive abilities, the general factor *g* is large and easy to spot. Specific variance is also easy to see in most tests. However, group factors typically explain only small amounts of variance in subtests, and this often makes group factors difficult to identify with confidence.

Clinicians are thus put in a tough spot. General ability scores are not only famously stable and reliable, but they are unmatched in terms of the number of studies providing evidence of predictive validity (Gottfredson, 1997). Yet general ability scores do not offer the kind of nuanced information clinicians hope to have so that they can understand the highs and lows of a person's performance on a wide variety of tasks across diverse settings. They want to understand why people can do well in some situations and on some tasks but have great difficulty in other situations and on other tasks. Group factors, in theory, offer some of the nuance clinicians need. Although the collection of supporting validity studies for group factors is not as impressive as the body of evidence supporting general ability scores, it is clear that group factors are differentially associated with important outcomes (McGrew & Wendling, 2010). Unfortunately, there is little evidence that clinicians are able to measure the non-*g* portions of group factors with precision, make valid inferences about them, and use this knowledge to help individuals (Canivez, 2013; Glutting, Watkins, Konold, & McDermott, 2006).

I note here that my beliefs about the theoretical construct *g* are more complex than they might appear in this presentation because some simplification was needed to make the discussion below clear. This article will do nothing to advance any theoretical position about *g*. When I refer to *g*, I am referring to the *g* implied by the models in the target articles, not necessarily to a theoretical causal entity independent of those models. I believe that given the evidence, no one can know with certainty what *g* is, if it is anything at all. My best guess is that *g* exists but that it is not really an ability. It is instead the result of many influences on the brain as a whole. For example, when lead poisoning occurs, it kills neurons more or less indiscriminately all over the brain, disrupting many cognitive functions simultaneously. It is possible that *g* emerges because there are many such factors that have a global influence on the brain (e.g., malnutrition, toxins, genetic abnormalities, blunt force trauma, large strokes, etc.). Even so, it makes sense to estimate *g* as an index of the overall integrity of the cerebral cortex. It is possible that even if *g* is not really an ability, for all practical purposes it might behave as if it were, making the distinction between *g* as *general intelligence* and *g* as *intelligence-in-general* a hair-splitting parlor game for theorists only.

Stumbling Blocks to Understanding How to Interpret Non-*g* Factors

Measures of Cognitive Abilities Other Than g Are Not Independent of g

In higher-order factor models, often most of what constitutes the lower-order factors is *g*. In the final five-factor WISC-IV model, for example, 71% of Gc's variance comes from *g*. According to this model, it can be shown that about 60% of the variance in the WISC-IV Verbal Comprehension Index (VCI) is due to *g*. Thus, it is important to remember that most measures of "Gc" contain a substantial amount of *g* variance in them, and thus, they are partially redundant if a measure of *g* has already been taken into account. Note that the independent portion is not the "real Gc." We care about a sprinter's ability to run quickly, not residual sprinting speed after accounting for general athleticism. So it is with Gc: *g* is a part of the mix. Even so, we still would like to estimate the portion of Gc that is independent of *g*.

Reliability Is Not the Same as Validity

When I discuss the problems of estimating abilities other than *g*, clinicians sometimes counter with evidence that their non-*g* ability measures are highly reliable. This is certainly true, but much of that reliability is due to *g* variance, and much of the rest is reliable but subtest-specific variance. Our ability to validly operationalize non-*g* factors is far worse than many clinicians suppose.

There is a tired old joke about the drunk who lost his keys on the dark side of the street but is looking for them under the lamppost because "That's where the light is." Reliability is where the light is. Validity is where the keys are. Reliability is relatively easy to estimate compared to validity. Researchers and test developers make a very big deal out of high reliability coefficients because "A test cannot be valid if it is not reliable." However, the fact that a measure is highly reliable is irrelevant if it does not allow us to make accurate inferences about the thing we wish to measure. Furthermore, if a measure is shown to have validity, its reliability is already implied. To switch metaphors, reliability is thin gruel if validity is on the table. I think that with good models such as those offered by the target articles, validity is at least on the menu, if not already laid out for the feast. Reliability is at best an appetizer. It is nice to have, but if the main course is ample, you can skip it without worries.

"Significant Differences" Are Not Very Meaningful

When a statistical test reveals that two scores are "significantly different" from each other, it merely means that the observed difference is unlikely to be due to measurement error alone. It means nothing about the size or clinical relevance of the difference. The true scores might have a trivial difference such as 1 or 2 index score points and this is counted as "significant" because the difference is not 0. It would be better to estimate the probability that the true scores are "meaningfully different." If I believe that a true score difference must be at least 10 points before it can be clinically meaningful, it would be more useful to calculate the probability that the true score difference is 10 points or more. Another problem of interpreting differences is that the two observed scores might differ for reasons other than those supposed. For example, the WISC-IV FSIQ (Full-Scale IQ) and VCI might differ not because of the non-*g* portion of Gc but because of specific factors associated with Vocabulary, Similarities, and Comprehension. What is needed is a means of showing that a score likely differs from *g* because of the latent variable of interest and not because of some other set of factors.

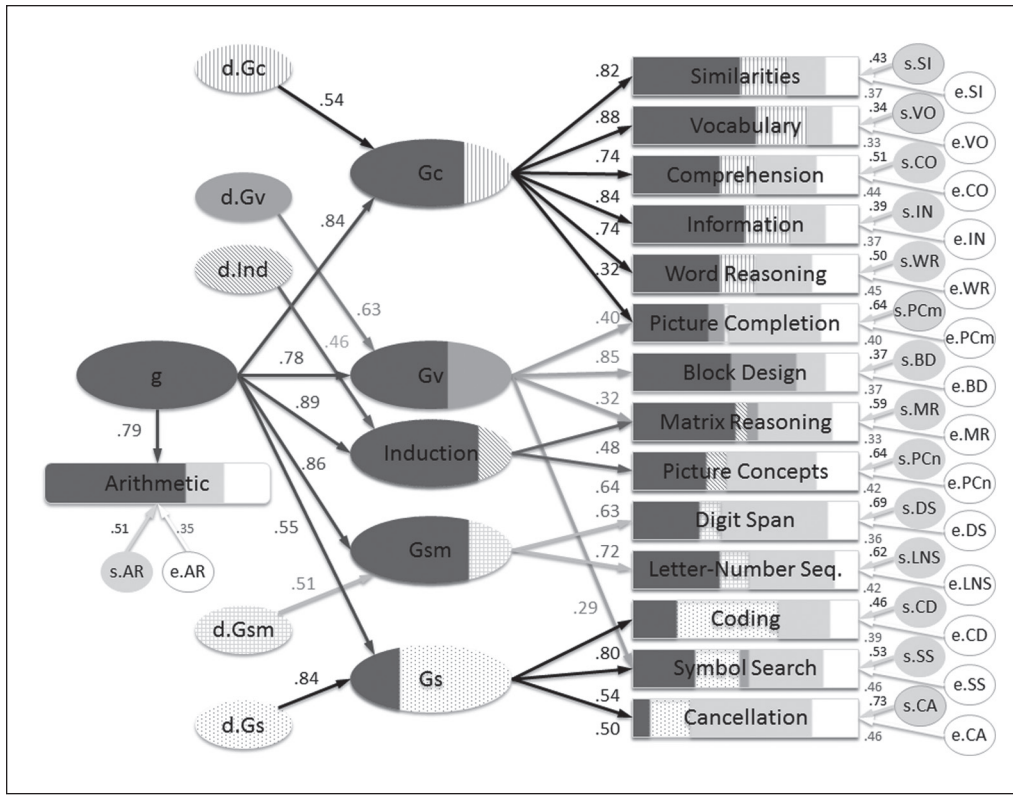

Figure 1. Modified WISC-IV (Wechsler Intelligence Scale for Children—Fourth Edition) five-factor model.

A Method for Estimating Latent Scores

The method I present here offers at least partial solutions to these problems. I will illustrate the method with the final five-factor model for the WISC-IV (Weiss et al., 2013b) but with a small modification that streamlines the calculations. In that model, *g* and Gf are perfectly correlated, meaning that Gf is redundant. As seen in Figure 1, eliminating Gf from the model makes Induction a direct effect of *g* and Arithmetic a direct indicator of *g*. In no way does this modification affect the interpretation of the model. Another difference in Figure 1 from the final five-factor model is that I have added the error terms, which have been estimated from the overall reliability coefficients from the standardization sample (Wechsler, 2003).

I am hardly the first scholar to note that it is possible to estimate latent scores from observed scores. Indeed, Thurstone (1935), the primary inventor of multiple factor analysis, also invented the regression method I use in this article for estimating factor scores. It is common for researchers to estimate latent variable scores so that they can be used in subsequent analyses (for a discussion of the various problems of doing so, see Grice, 2001). Oh, Glutting, Watkins, Youngstrom, and McDermott (2004) presented a method of estimating latent scores that is designed for clinicians. The methods I present here extend the ideas of Oh and colleagues in several ways. First, it is important not only to estimate latent scores but also calculate appropriate confidence intervals around them. Second, it is useful not only to estimate exogenous and endogenous latent variable scores but also disturbance and specific scores. Disturbances are often mere residuals. However, the disturbance score for Gc (d.Gc in Figure 1) represents the part of Gc that is independent of *g*. If Gc were perfectly predicted from *g*, there would be no point in estimating it. Far from being irrelevant, d.Gc has real theoretical heft.

In Figure 1, each subtest has two "disturbances," one representing reliable specific influences and the other representing measurement error. Together, specific variance and error variance compose variance that makes each subtest score unique (uniqueness = specific + error). Specific and error true scores are always uncorrelated. Although it is possible to estimate both the specific score and the error score for a subtest, those estimates will unfortunately be perfectly correlated (but with different standard deviations). That is, we can estimate latent uniqueness scores, but there is no way to make the estimated specific scores and estimated error scores independent of each other. Fortunately, we are typically not interested in estimating error scores. They are modeled solely because they put limits on the accuracy of the estimated specific scores.

Specifying the Model

From the observed subtest scores, we would like to estimate the latent exogenous variable scores (in this case, *g*), the latent endogenous variable scores (Gc, Gv, Induction, Gsm, and Gs), the disturbance scores associated with the latent endogenous variables (d.Gc and so forth), the specific scores associated with each subtest (s.SI and so forth), and the error scores associated with each subtest (e.SI and so forth). To do this, we will need model parameter estimates arranged in three matrices: Φ, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$.

From these matrices, we will be able to calculate model-implied correlations among all the variables. From these model-implied correlations and the observed correlations between the subtests, we will be able to construct regression equations that estimate latent scores. In all calculations we will assume that all variables are multivariate normal. We will also assume that the model is correct and that all of the parameters in the model have been estimated perfectly. This is the where the title of this article comes from: We are imagining what could be inferred from observed test scores if we take the model seriously and assume that it is correct.

In a sense, disturbance variables, specific influences, and measurement error are exogenous latent variables. That is, their effects are determined by forces outside the model. Thus, it makes sense to group all the exogenous variables together (i.e., the latent exogenous variables, observed exogenous variables [absent from this model], disturbance variables, specific influences, and errors) in one vector we can call *xi* ($\boldsymbol{\xi}$). All notation and most of the following formulas come directly or indirectly from the Bentler–Weeks model (Bentler & Weeks, 1980). The matrix of covariances among all the variables in $\boldsymbol{\xi}$ is called *phi* ($\boldsymbol{\phi}$). To make things simple, we will assume that all of the variables in ξ are standardized ($M = 0$, $SD = 1$). This means that all of the elements in $\boldsymbol{\phi}$ are correlations instead of covariances. In this example, none of the exogenous variables are correlated, but many models have correlations among error terms and among other exogenous variables.

It also makes sense to think of observed indicators of latent variables as observed endogenous variables because they are fully determined by the exogenous variables in $\boldsymbol{\xi}$. The vector of endogenous variables can be called *eta* ($\boldsymbol{\eta}$). Thus, it makes sense to place the standardized effects of all 21 endogenous variables (both latent and observed) on other endogenous variables in a 21×21 matrix called *beta* ($\boldsymbol{\beta}$). The standardized effects of the 36 exogenous variables in $\boldsymbol{\xi}$ on the 21 endogenous variables in $\boldsymbol{\eta}$ are placed in the 36×21 matrix called *gamma* ($\boldsymbol{\gamma}$).

Calculations

The endogenous variables can be computed from the exogenous variables like so:

$$\boldsymbol{\eta} = (\mathbf{I} - \boldsymbol{\beta})^{-1} \boldsymbol{\gamma}\boldsymbol{\xi}$$

However, we would like an equation in which all of the variables in $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are on the left side. This is accomplished by appending a compatible identity matrix to $\boldsymbol{\gamma}$ and some compatible matrices of zeros to $\boldsymbol{\beta}$ like so:

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\gamma} \\ \mathbf{I} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

The vectors $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are appended to create a vector *nu* ($\boldsymbol{\nu}$) of all of the variables:

$$\boldsymbol{\nu} = \begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\xi} \end{pmatrix}$$

Thus,

$$\boldsymbol{\nu} = (\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}\boldsymbol{\xi}$$

The model-implied correlations of all the variables in ν are calculated like so:

$$\mathbf{R}_{\nu} = (\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi}((\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma})'$$

Once the correlations among all the variables have been calculated, a set of regression equations can be calculated to estimate the latent variable scores from the observed scores. Standardized regression coefficients for estimating scores for the latent variables are calculated like so:

$$\boldsymbol{\beta}_{Subtest \times Latent} = \mathbf{R}^{-1}_{Subtest \times Subtest}\mathbf{R}_{Subtest \times Latent}$$

Where:

$\boldsymbol{\beta}_{Subtest \times Latent}$ is the matrix of standardized regression coefficients (where each column represents the coefficients for a regression equation for estimating a single latent variable).

$\mathbf{R}_{Subtest \times Subtest}$ is the matrix of observed correlations among all the subtests.

$\mathbf{R}_{Subtest \times Latent}$ is the matrix of model-implied correlations between each latent variable and each subtest.

The coefficient of determination (multiple R^2) associated with each regression equation can be calculated like so:

$$\mathbf{d} = \text{diag}(\boldsymbol{\beta}'_{Subtest \times Latent}\mathbf{R}_{Subtest \times Latent})$$

The standardized regression coefficients for calculating estimated latent scores in this model are displayed in Table 1.

The standard error of the estimate is used to create confidence intervals around estimated latent scores. The standard error of the estimate for each equation can be calculated simultaneously like so:

$$\boldsymbol{\sigma}_e = \sqrt{\mathbf{1} - \mathbf{d}}$$

Note that all latent scores are assumed to be *z*-scores. Also, the radical in the formula is an operation performed element-wise.

If, along with some well-placed zeros, the regression coefficients and a matrix of weights to calculate the FSIQ and the four-factor index scores are appended to the $\boldsymbol{\beta}$ matrix,

Table 1. Regression Coefficients for Predicting Latent Variables in the Modified Five-Factor WISC (Wechsler Intelligence Scale for Children) Model.

	SI	VC	CO	IN	WR	PCm	BD	MR	PCn	DS	LN	CD	SS	CA	AR	R^2
g	0.067	0.105	0.044	0.077	0.044	0.042	0.116	0.125	0.101	0.083	0.118	0.050	0.047	0.015	0.281	0.87
Gc	0.194	0.302	0.127	0.221	0.127	0.047	0.015	0.023	0.020	0.017	0.024	0.011	0.008	0.003	0.056	0.92
d.Gc	0.253	0.395	0.166	0.289	0.166	0.022	−0.153	−0.151	−0.118	−0.097	−0.139	−0.058	−0.058	−0.017	−0.331	0.57
Gv	0.012	0.019	0.008	0.014	0.008	0.132	0.562	0.147	0.018	0.023	0.034	−0.015	0.101	−0.005	0.080	0.82
d.Gv	−0.064	−0.100	−0.042	−0.073	−0.042	0.158	0.754	0.079	−0.097	−0.066	−0.094	−0.087	0.102	−0.026	−0.223	0.47
Ind	0.047	0.074	0.031	0.054	0.031	0.022	0.051	0.257	0.251	0.058	0.082	0.036	0.028	0.011	0.195	0.77
d.Ind	−0.028	−0.043	−0.018	−0.032	−0.018	−0.033	−0.115	0.319	0.354	−0.035	−0.050	−0.018	−0.031	−0.005	−0.119	0.17
Gsm	0.040	0.062	0.026	0.045	0.026	0.025	0.069	0.074	0.060	0.236	0.338	0.030	0.028	0.009	0.166	0.77
d.Gsm	−0.035	−0.055	−0.023	−0.040	−0.023	−0.022	−0.061	−0.066	−0.053	0.324	0.463	−0.026	−0.025	−0.008	−0.147	0.27
Gs	0.012	0.018	0.008	0.014	0.008	−0.002	−0.021	0.012	0.018	0.014	0.020	0.547	0.267	0.164	0.047	0.75
d.Gs	−0.030	−0.047	−0.020	−0.034	−0.020	−0.030	−0.102	−0.067	−0.045	−0.038	−0.054	0.623	0.289	0.187	−0.128	0.62
s.SI	1.112	−0.328	−0.137	−0.240	−0.137	−0.051	−0.016	−0.025	−0.022	−0.018	−0.026	−0.011	−0.009	−0.003	−0.061	0.48
s.VC	−0.257	1.106	−0.168	−0.293	−0.168	−0.062	−0.020	−0.031	−0.027	−0.022	−0.032	−0.014	−0.011	−0.004	−0.075	0.38
s.CO	−0.163	−0.253	1.026	−0.185	−0.106	−0.039	−0.012	−0.020	−0.017	−0.014	−0.020	−0.009	−0.007	−0.003	−0.047	0.53
s.IN	−0.217	−0.339	−0.142	1.087	−0.142	−0.053	−0.017	−0.026	−0.023	−0.019	−0.027	−0.012	−0.009	−0.004	−0.063	0.43
s.WR	−0.159	−0.248	−0.104	−0.182	1.006	−0.039	−0.012	−0.019	−0.017	−0.014	−0.020	−0.009	−0.007	−0.003	−0.046	0.51
s.PCm	−0.075	−0.117	−0.049	−0.086	−0.049	1.047	−0.258	−0.074	−0.015	−0.017	−0.024	0.003	−0.048	0.001	−0.056	0.67
s.BD	−0.014	−0.021	−0.009	−0.016	−0.009	−0.149	0.698	−0.167	−0.021	−0.027	−0.038	0.018	−0.115	0.005	−0.091	0.26
s.MR	−0.034	−0.053	−0.022	−0.039	−0.022	−0.068	−0.264	1.072	−0.163	−0.045	−0.065	−0.016	−0.059	−0.005	−0.154	0.63
s.PCn	−0.033	−0.051	−0.021	−0.037	−0.021	−0.015	−0.035	−0.178	0.911	−0.040	−0.057	−0.025	−0.019	−0.008	−0.136	0.58
s.DS	−0.029	−0.045	−0.019	−0.033	−0.019	−0.018	−0.049	−0.053	−0.043	0.971	−0.243	−0.021	−0.020	−0.006	−0.120	0.67
s.LN	−0.037	−0.057	−0.024	−0.042	−0.024	−0.023	−0.064	−0.068	−0.055	−0.218	0.970	−0.027	−0.026	−0.008	−0.154	0.60
s.CD	−0.012	−0.019	−0.008	−0.014	−0.008	0.002	0.022	−0.013	−0.018	−0.014	−0.020	0.715	−0.272	−0.167	−0.048	0.33
s.SS	−0.011	−0.017	−0.007	−0.012	−0.007	−0.040	−0.164	−0.053	−0.016	−0.015	−0.022	−0.314	0.892	−0.094	−0.053	0.47
s.CA	−0.006	−0.009	−0.004	−0.007	−0.004	0.001	0.010	−0.006	−0.009	−0.007	−0.010	−0.268	−0.131	0.899	−0.023	0.66
s.AR	−0.071	−0.111	−0.047	−0.081	−0.047	−0.044	−0.124	−0.133	−0.107	−0.088	−0.126	−0.053	−0.050	−0.016	1.047	0.53

Table 2. Correlation of Latent Variables With Their Respective Estimated Latent Variables.

Variable	*r*
g	0.93
Gc	0.96
d.Gc	0.75
Gv	0.90
d.Gv	0.69
Induction	0.88
d.Induction	0.42
Gsm	0.88
d.Gsm	0.52
Gs	0.87
d.Gs	0.79
s.SI	0.69
s.VC	0.61
s.CO	0.72
s.IN	0.65
s.WR	0.71
s.PCm	0.82
s.BD	0.51
s.MR	0.79
s.PCn	0.76
s.DS	0.82
s.LN	0.77
s.CD	0.57
s.SS	0.69
s.CA	0.81
s.AR	0.73
FSIQ and *g*	0.90
VCI and Gc	0.92
PRI and Gv	0.81
PRI and Induction	0.80
WMI and Gsm	0.79
PSI and Gs	0.84

the model-implied correlations among all the observed variables, the latent variables, and the estimated latent variables can be calculated using the same procedure as before. This time, however, the new variables will not be *z*-scores, and it is necessary to convert the covariance matrix to a correlation matrix.

Because all the exogenous variables (*g*, errors, and specifics) are uncorrelated, it is possible to determine from the correlation matrix how much each observed variable and how much each estimated latent variable is influenced by the latent variables. As can be seen in Table 2, the estimated latent scores are better predictors of the latent constructs than are the FSIQ and the four-factor index scores. In part, the estimated latent scores perform better because they use 15 subtests instead of just 10. However, because the subtests are optimally weighted (assuming the model is correct), the estimated latent scores are more accurate than the composite scores even if restricted to the 10 subtests used by the FSIQ. Figure 2 shows the variance composition of the composite scores and the estimated latent scores.

Table 3. WISC-IV Scores for Hypothetical Case.

IQ/index/subtest	Score
Full-Scale IQ	101
Verbal Comprehension Index	114
Similarities	13
Vocabulary	15
Comprehension	10
Information	8
Word reasoning	11
Perceptual Reasoning Index	108
Block design	9
Picture concepts	12
Matrix reasoning	13
Picture completion	12
Working Memory Index	94
Digit span	9
Letter–number sequencing	9
Arithmetic	7
Processing Speed Index	75
Coding	5
Symbol search	6
Cancellation	7

Note: WISC-IV = Wechsler Intelligence Scale for Children—Fourth Edition.

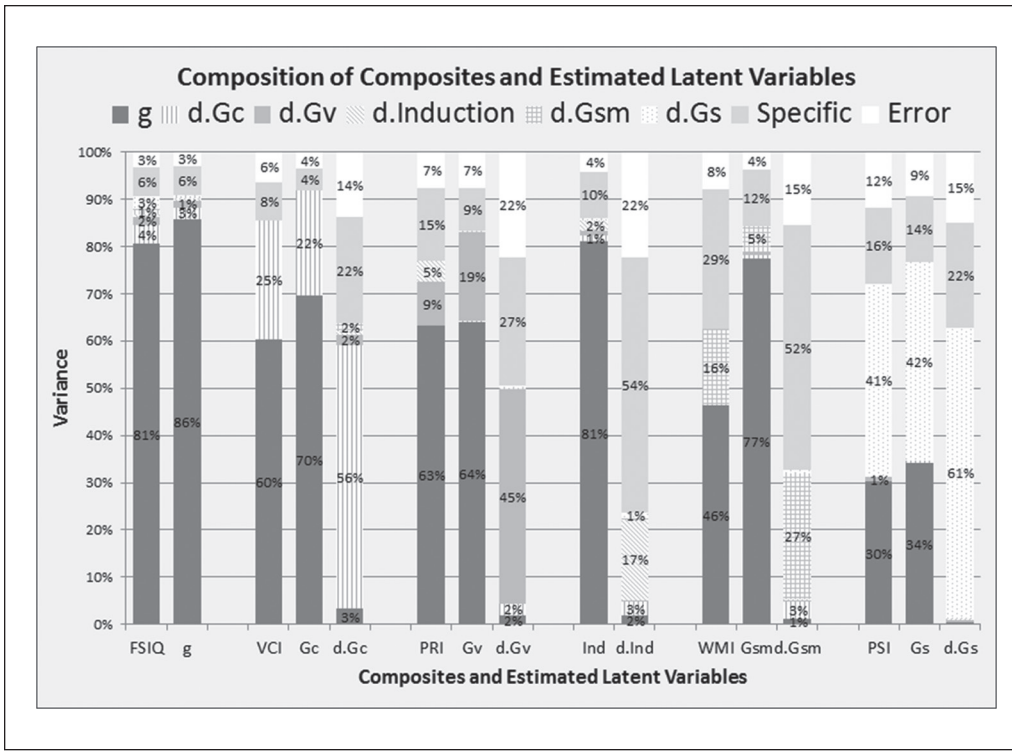

Figure 2. Composition of composite scores and estimated latent variables.

FSIQ Versus Estimated *g*

It is noteworthy that FSIQ has a correlation of "only" .90 with *g*. It is customary to regard FSIQ as an operationalization of *g*, and it is easy (but incorrect) to think of the reliability coefficient of FSIQ (.97) as its correlation with *g*. Reliability refers to the percentage of true score variance in the observed score. The square root of the reliability coefficient is the observed score's correlation with the true score but the true score is not synonymous with *g*. In this case, as seen in Figure 2, the "true score" associated with FSIQ is an amalgamation of *g* (81%), d.Gc, d.Gv. d.Induction, d.Gsm, d.Gs (10% all combined), and 10 reliable but specific subtest influences (6% all combined). Thus, FSIQ is about 97% reliable, but only 81% of its variance is caused by *g*. About 16% of FSIQ's variance is reliable but unrelated to *g*. Some of this non-*g* variance undoubtedly contributes to FSIQ's predictive validity in many applications. If the purpose of FSIQ is to predict outcomes, this non-*g* variance is "valid." However, if FSIQ is used to operationalize *g*, then its validity is lower than many people assume.

The estimated *g* score has more *g* variance in it than does FSIQ (87% vs. 81%). This means that when *g* is correlated with an outcome and the other factors are not correlated with it, the estimated *g* score will have greater predictive validity than will FSIQ. If however, the non-*g* factors in FSIQ are also correlated with the outcome, FSIQ might have greater predictive validity than will the estimated *g* score. Fortunately, those other factors can be estimated as well.

Estimating Scores and Confidence Intervals

If an individual's subtest scores are converted to *z*-scores and arranged in a row vector $\boldsymbol{s}$, the estimated latent scores are calculated like so:

$$\hat{\boldsymbol{v}} = \boldsymbol{s}\boldsymbol{\beta}_{Subtest\ x\ Latent}$$

The 95% confidence intervals (95% CI) for the latent scores are obtained by multiplying the standard error of the estimate by the appropriate *z*-score:

$$\boldsymbol{v} = \hat{\boldsymbol{v}} \pm z_{95\%}\boldsymbol{\sigma}_e$$

We also want to make specific predictions about whether a latent score is above or below a threshold a clinician might consider meaningful. For example, we might want to estimate the probability that, given an estimated *g* score, the true latent *g* score is 1 *SD* or more below the mean. These conditional probability estimates are calculated like so:

$$p(\boldsymbol{v} \le \mathbf{z}|\hat{\boldsymbol{v}}) = N\left(\frac{\mathbf{z} - \hat{\boldsymbol{v}}}{\boldsymbol{\sigma}_e}\right)$$

where

$p(\boldsymbol{v} \le \mathbf{z}|\hat{\boldsymbol{v}})$ is a vector of probabilities that the true latent scores $\boldsymbol{v}$ are less than or equal to a compatible vector of thresholds in $\mathbf{z}$, given the estimated latent scores in $\hat{\boldsymbol{v}}$.
$N(x)$ is the cumulative density function of the standard normal distribution (i.e., the proportion of the standard normal curve that is less than x).
$\hat{\boldsymbol{v}}$ is the vector of estimated latent scores.
$\boldsymbol{\sigma}_e$ is the vector of standard errors of the estimate when observed subtest scores are used to predict latent scores in $\boldsymbol{v}$.

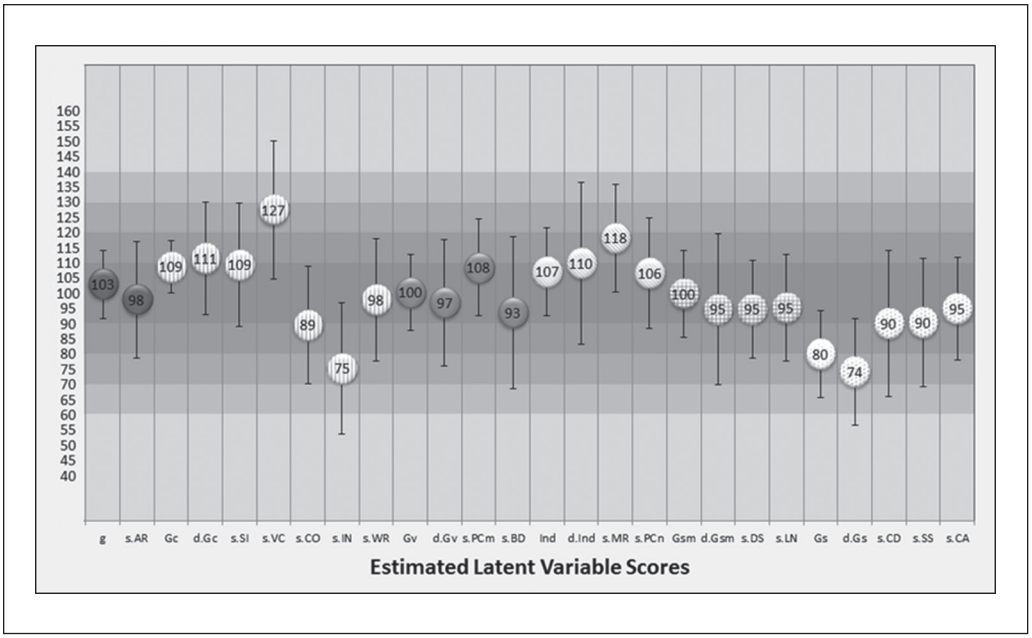

Figure 3. Hypothetical case profile of estimated latent variable scores with 95% confidence intervals.

Test Interpretation With a Hypothetical Case

Suppose that a child has a WISC-IV profile as shown in Table 3. The spreadsheet program that accompanies this article (available from http://my.ilstu.edu/~wjschne/tests.html) uses the equations from above to calculate the estimated latent scores and their confidence intervals, as shown in Figure 3. For ease of interpretation, they are shown in the index score metric ($M = 100$, $SD = 15$). In addition, the probability that the true latent scores are above or below some selected thresholds is displayed in Figure 4.

Noteworthy Features of Estimated Latent Scores

First, estimated latent scores are not the true latent scores. They are estimates with a certain amount of error and construct-irrelevant influences. Some estimates are better than others. Second, although they appear to be on the same scale as the true latent scores, their variance is equal to the variance of the true latent scores multiplied by the multiple R^2 from the regression equation between the subtest scores and the latent scores. That is, they have less variance than the true latent scores. In Figure 4, the reliable specific influence on Similarities (s.SI) is estimated to be 110. On the index score metric, this corresponds to the 75th percentile. However, this does not mean that 25% of children have an estimated score of 110 or higher. Actually only about 17% of children are estimated to score at the 75th percentile or higher. Counterintuitive? Yes. Does it have something to do with regression to the mean? Yes, most counterintuitive ideas in statistics do.

Think of it this way: If the estimated latent score had a correlation of zero with the true latent score, our best estimate would be that everyone's true score was at the mean, or the 50th percentile, and thus the estimated score would have no variance at all. When our predictions are uncertain, we must be conservative. If the correlation were positive but small, we would use the regression formula, but the regression toward the mean would be severe and we would still guess

Estimated Scores		95% CI	Probability that the true latent score is: <60	<70	<80	<90	<100	>100	>110	>120	>130	>140
g	103	92 to 114	0	0	0.00003	0.01	0.31	0.69	0.11	0.001	0	0
s.AR	98	79 to 117	0.00006	0.002	0.03	0.21	0.59	0.41	0.11	0.01	0.0005	0.00001
Gc	109	100 to 117	0	0	0	0.00001	0.03	0.97	0.37	0.005	0	0
d.Gc	111	93 to 130	0	0.00001	0.0005	0.01	0.12	0.88	0.56	0.18	0.02	0.001
s.SI	109	89 to 130	0	0.00007	0.002	0.03	0.18	0.82	0.47	0.15	0.02	0.002
s.VC	127	105 to 150	0	0	0.00002	0.0007	0.009	0.991	0.93	0.74	0.41	0.14
s.CO	89	70 to 109	0.001	0.02	0.17	0.52	0.86	0.14	0.02	0.001	0.00002	0
s.IN	75	53 to 97	0.09	0.32	0.67	0.91	0.99	0.01	0.0008	0.00003	0	0
s.WR	98	78 to 118	0.0001	0.004	0.04	0.23	0.59	0.41	0.12	0.02	0.0009	0.00002
Gv	100	88 to 113	0	0	0.0007	0.05	0.49	0.51	0.06	0.0009	0	0
d.Gv	97	76 to 117	0.0003	0.006	0.06	0.26	0.62	0.38	0.1	0.01	0.0008	0.00002
s.PCm	108	92 to 124	0	0	0.0002	0.01	0.15	0.85	0.42	0.08	0.004	0.00005
s.BD	93	68 to 119	0.004	0.03	0.15	0.39	0.69	0.31	0.1	0.02	0.002	0.0001
Induction	107	93 to 121	0	0	0.0001	0.01	0.17	0.83	0.34	0.04	0.0008	0
d.Induction	110	83 to 136	0.0001	0.002	0.01	0.07	0.24	0.76	0.49	0.22	0.07	0.01
s.MR	118	100 to 136	0	0	0.00001	0.0009	0.02	0.98	0.82	0.42	0.09	0.007
s.PCn	106	88 to 125	0	0.00005	0.002	0.04	0.24	0.76	0.35	0.07	0.006	0.0002
Gsm	100	85 to 114	0	0.00002	0.003	0.09	0.51	0.49	0.08	0.003	0.00002	0
d.Gsm	95	70 to 120	0.003	0.03	0.12	0.36	0.66	0.34	0.11	0.02	0.003	0.0002
s.DS	95	79 to 111	0.00001	0.001	0.04	0.28	0.74	0.26	0.03	0.001	0.00001	0
s.LN	95	78 to 113	0.00004	0.002	0.04	0.28	0.71	0.29	0.05	0.003	0.00005	0
Gs	80	66 to 94	0.003	0.09	0.5	0.92	0.997	0.003	0.00002	0	0	0
d.Gs	74	57 to 91	0.06	0.32	0.75	0.96	0.998	0.002	0.00003	0	0	0
s.CD	90	66 to 114	0.007	0.05	0.21	0.5	0.8	0.2	0.05	0.007	0.0005	0.00002
s.SS	90	69 to 111	0.002	0.03	0.17	0.49	0.82	0.18	0.03	0.003	0.0001	0
s.CA	95	78 to 112	0.00003	0.002	0.04	0.29	0.72	0.28	0.04	0.002	0.00003	0

Figure 4. Probabilities associated with the hypothetical child's true latent scores being under or over selected thresholds.

that everyone fell near the 50th percentile. Thus, the variance of the estimated score would be very small. As the correlation between the estimate and the true latent score increases, the regression toward the mean becomes less and less severe and the estimated latent variable's standard deviation approaches that of the true latent variable.

Two Kinds of Confidence Intervals

If the 95% CI of the WISC-IV FSIQ is calculated from the test manual, the confidence interval is about 10 index score points wide (in this case from 96 to 106). Although both FSIQ and estimated *g* have a reliability coefficient of .97, the confidence interval for estimated *g* in Figure 4 is about 22 points wide (in this case from 90 to 112). What is going on here? Shouldn't they be nearly the same? Yes, except that these are not the same kind of confidence intervals. The first is a reliability-based confidence interval. The second is a validity-based confidence interval. The reliability-based confidence interval estimates the location of the true score (the sum of all reliable influences). The validity-based confidence interval estimates the location of the latent variable that the test is intended to operationalize. If FSIQ's purpose were solely to estimate *g*, its validity-based confidence interval would be 26 points wide. Thus, the estimated *g* score does a better job of locating a person's level of *g* than does FSIQ. In the case of this particular person, *g* is likely to be somewhere in the average range but whether it is at the upper end of average, the lower end, or, more likely, right in the middle, is impossible to know.

Interpreting Gc

At 16 points wide, the 95% CI for estimated Gc is even narrower than the 95% CI for estimated *g*. This might be counterintuitive because we are used to thinking of factor scores as being less reliable than general ability scores. However, the estimated *g* score and the estimated Gc score are both estimated from all 15 subtest scores (but use different weights). In truth, the estimated *g* score is a bit more reliable than the estimated Gc score, but it does not matter: The validity coefficient for estimated Gc (.96) is stronger than the validity coefficient for estimated *g* (.93). A stronger validity coefficient makes the validity-based confidence interval narrower.

As can be seen in Figures 3 and 4, this person's Gc is likely to be on the upper end of average. There is a 34% chance that Gc is above 110 (substantially above average), but there is less than a 1% chance that it is above 120. The reason that estimated Gc is a little higher than estimated *g* is that there is a 93% probability that the non-*g* portion of Gc (d.Gc) is above the mean. The confidence interval for d.Gc is much wider than the confidence interval for Gc (38 vs. 16 points). Because the 95% CI for d.Gc (96 to 134) contains the population mean of 100, *g* and Gc cannot be said to be "significantly different" from one another (although the odds are not bad that they are in fact different).

Normally the specific factors associated with subtest scores are not of theoretical interest. However, a reasonable case can be made for looking at them when a specific factor is likely to correspond to a known theoretical construct that correlates with important outcomes. For example, the Vocabulary subtest measures Carroll's (1993) Lexical Knowledge factor in addition to *g* and Gc. We can imagine that if the WISC-IV had another measure of vocabulary such as a picture vocabulary test, a portion of Vocabulary's specific variance would form a narrow Lexical Knowledge factor. Because Lexical Knowledge plays such an important role in so many academic and occupational outcomes, it is reasonable to look at the specific factor associated with the Vocabulary subtest.

The s.VC score is not the person's Lexical Knowledge ability in its totality. Rather, the s.VC score is the influence of a specific talent (or weakness) in Lexical Knowledge after accounting for the effects of *g* and Gc. Lexical Knowledge as a whole is better estimated by the observed Vocabulary score. However, we can estimate how much different abilities contributed to the final observed score, as shown in Figure 5. The estimated effect of *g* starts the score at nearly the mean. The estimated effect of d.Gc adds about half a standard deviation to the score. d.Gv, d.Induction, d.Gsm, and d.Gs have no effect at all (according to the model). The specific effect of Vocabulary (s.VC) and the error effect (e.VC) are estimated to add almost 0.6 *SD* each such that the final observed score is 15 (1⅔ *SD* above the mean). Note that these estimates only show the most probable pattern of influences and that the truth might be quite different. Even so, Figure 5 might assist the clinician's intuition about how the Vocabulary subtest score of 15 might have come about. The spreadsheet can provide a similar chart for all of the subtests, FSIQ, and the four-factor index scores. It should be kept in mind that the latent variable estimates are not independent of each other. If somehow *g* were known to be lower than estimated *g*, all of the other estimates would have to be updated to reproduce the observed scores.

Just as it appears that s.VC is significantly higher than average, s.IN (the reliable specific influence on the Information subtest) is at least a bit lower than average. That is, compared to other aspects of Gc, Lexical Knowledge may be a bit better developed and the narrow factor of General Knowledge lags a bit behind other aspects of Gc.

This practice of examining specific effects on single subtests runs counter to frequently heard injunctions to NEVER interpret single subtest scores because they are "not reliable enough." However, the broader principle is not to make interpretations from unreliable data. The confidence intervals around the specific scores take the low reliability of the single scores into account. Thus, if the score deviates significantly from the mean, it is permissible to interpret it. Obviously

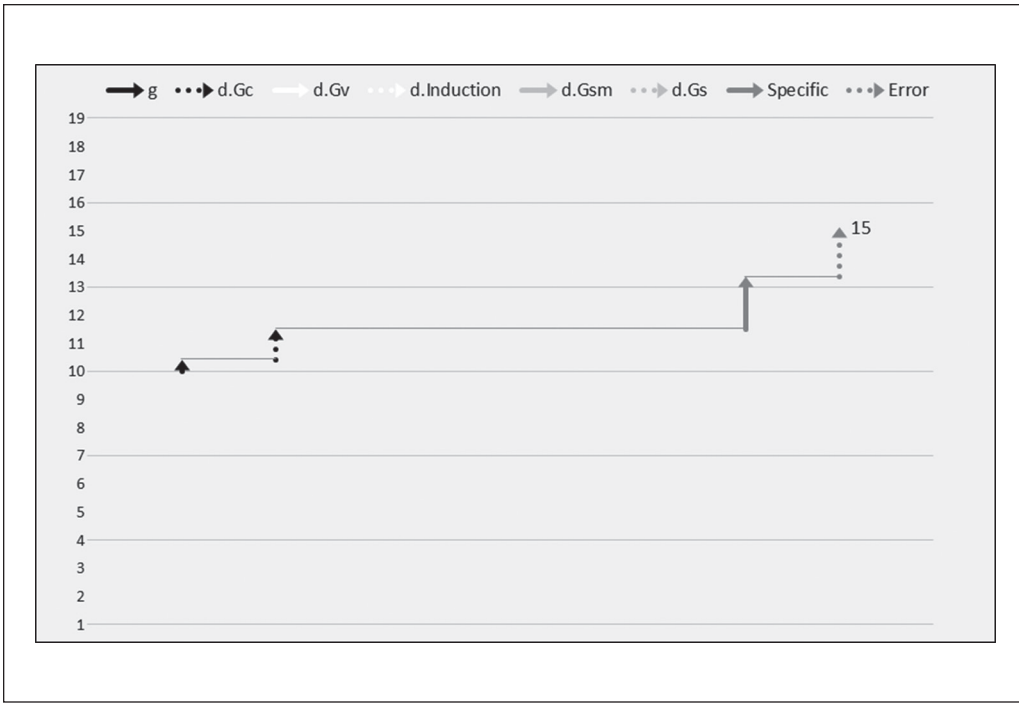

Figure 5. Estimated influences on the hypothetical child's Vocabulary subtest score.

follow-up testing with a similar test is desirable to confirm findings about test-specific influences. One should also keep in mind that this model does not address the long-term stability of test-specific influences and thus an extra dollop of caution is warranted.

Interpreting Gv, Induction, and Gsm

In this child's profile, the estimated Gv score is not significantly different from the estimated *g* score (i.e., the estimated d.Gv score's confidence interval contains the mean). This is also true of the estimated Induction score and the estimated Gsm score. The estimated d.Induction score has a confidence interval that is 53 points wide, which is about 3½ *SD*. The estimated d.Induction variable has a correlation of .42 with the true d.Induction latent variable. A correlation of .42 might be a pretty substantial relationship in many applications. However, for the purpose of estimating individual scores, it is very low. In fact, if the estimated d. Induction score were completely uncorrelated with the true d.Induction latent variable, its "95% confidence interval" would be 59 points wide. The fact that the d.Induction confidence interval is only a little smaller than 59 points means that very little can be said with confidence about the difference between Induction and *g*, except in extreme cases. Future editions of the WISC will probably have to include more than two subtests that measure inductive reasoning if clinicians are to have any hope of measuring this ability with precision and distinguishing it from *g*.

Interpreting Gs

Of all the group factors measured by the WISC-IV, Gs is the most independent of *g*. Only 30% of its variance is due to *g*. This means that Gs is the group factor that will deviate from *g* most often. The true latent Gs score is also well predicted by the estimated Gs score ($r = .87$). This

means that it is also easier to be confident about differences between Gs and *g* when they occur. In the case of this child, it is clear that the estimated Gs of 79 (95% CI = 64-94) is lower than the mean. Estimated d.Gs is 74 (95% CI = 56-92). Because the confidence interval of d.Gs does not contain the mean (100), it is likely that this child's true Gs score is lower than the true *g* score. However, just because Gs and *g* are "significantly different" from each other in the statistical sense, it does not necessarily mean that the difference is substantial or clinically important. Indeed, the confidence interval for Gs extends all the way up to 94. In Figure 4, it can be seen that Gs is very likely (99.8%) to be less than the average of 100, pretty likely (93%) to be less than 90, the lower end of the average range, and is fairly likely (54%) to be less than 80, the lower end of the low average range. I imagine that most clinicians would interpret this child's PSI (Parenting Stress Index) of 75 as strong evidence that Gs is low. It very well may be quite low, but this conclusion is far less certain than it would appear if one were merely examining the reliability-based confidence interval of the PSI.

Case Summary

This child likely has average *g* with Gc in the average to high average range. It is likely that this child has at least a somewhat better developed vocabulary than other aspects of Gc and might have a somewhat underdeveloped fund of general knowledge. It is hard to say anything precise about Gv, Gsm, and Induction other than that it is unlikely that they are extremely high or extremely low. Most likely they are average. It is likely that Gs is at best low average, and it is fairly likely that it is a good deal lower than that. It is therefore probable that Gs is low enough to be clinically meaningful, but it is impossible to be confident that this is so. Follow-up testing with more measures of Gs would clarify the matter considerably.

Advantages of This Method

What about this interpretation is different from one that relies solely on observed scores? I would argue two things: clarity and humility. That is, in terms of the overall description of the shape of the profile, there is little difference between this approach and more traditional ones. However, the clinician can be much clearer about how uncertain the estimates of *g* and other abilities are. Why is this desirable? Why would clinicians want methods that make them feel less certain about what their test scores mean? Having one's confidence stripped away is not anyone's idea of a good time. But it is not enough to be *confident* that one is right. One must actually *be* right. I do not for a second believe that this model is "right" in the absolute sense, but I believe that it helps clinicians arrive at closer approximations of the truth, even if the truth is that some things can only be known approximately.

Thus, with apologies to Winston Churchill, I have nothing to offer but doubt, dread, fuss, and strain. You ask, what is our aim? I can answer in one word: Validity. Validity at all costs—Validity in spite of all illusions—Validity, however daunting or dreary the task may be, for without validity there is no purpose.

Future Directions

An obvious next step is to develop a consensus structural model of the relations among the Wechsler scales and the Wechsler Individual Achievement Test–Third Edition (WIAT-III; or between cognitive abilities and academic achievement variables more generally). From there, a more rigorous and precise definition of specific learning disorders can be developed. Clinicians could estimate how much of a person's academic difficulties is due to *g* and how much of it is due to non-*g* portions of other abilities. If the non-*g* portions of other abilities are estimated to

negatively influence academic abilities to a clinically meaningful degree (e.g., more than 1 *SD*), a strong case can be made that the person has a specific learning disorder.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

References

Bentler, P. M., & Weeks, D. G. (1980). Linear structural equations with latent variables. *Psychometrika*, *45*, 289-308.

Canivez, G. L. (2013). Psychometric versus actuarial interpretation of intelligence and related aptitude batteries. In D. H. Saklofske, & V. L. Schwean (Eds.), *Oxford handbook of psychological assessment of children and adolescents* (pp. 84-112). New York, NY: Oxford University Press.

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

Glutting, J. J., Watkins, M. W., Konold, T. R., & McDermott, P. A. (2006). Distinctions without a difference: The utility of observed versus latent factors from the WISC-IV in estimating reading and math achievement on the WIAT-II. *Journal of Special Education*, *40*, 103-114.

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence*, *24*, 79-132.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*, 430-450.

McGrew, K. S., & Wendling, B. J. (2010). Cattell–Horn–Carroll cognitive–achievement relations: What we have learned from the past 20 years of research. *Psychology in the Schools*, *47*, 651-675.

Oh, H. J., Glutting, J. J., Watkins, M. W., Youngstrom, E. A., & McDermott, P. A. (2004). Correct interpretation of latent versus observed abilities: Implications from structural equation modeling applied to the WISC-III and WIAT linking sample. *Journal of Special Education*, *38*, 159-173.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology*, *15*, 201-293.

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

Thurstone, L. L. (1935). *The vectors of the mind*. Chicago, IL: University of Chicago Press.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children—Fourth Edition*. San Antonio, TX: Harcourt Assessment, Inc.

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013a). WAIS-IV and clinical validation of the four- and five-factor interpretative approaches. *Journal of Psychoeducational Assessment*, *31*, 94-113.

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013b). WISC-IV and clinical validation of the four- and five-factor interpretative approaches. *Journal of Psychoeducational Assessment*, *31*, 114-131.