The Wechsler Intelligence Scale for Children III and the Cognitive Abilities Test (Form 6): Are the general factors the same?

David F. Lohman[‡]
University of Iowa
July 1, 2003

This study investigated the concurrent validity of the Wechsler Intelligence Scale for Children—Third Edition (WISC-III; Wechsler, 1991) and Form 6 of the Cognitive Abilities Test (CogAT6; Lohman & Hagen, 2001a). The standard battery of WISC-III and level D of CogAT6 were administered to 91 6th-grade students. The representativeness of the sample was first investigated by comparing the covariance matrix for each test battery with the covariance matrix reported for the standardization sample. Relationships between observed scores and latent factors on the two test batteries were then explored. It was expected that the general and verbal factors on the two tests would be highly correlated and that the WISC-III Performance scale would show highest correlations with the CogAT6 Nonverbal score. Latent factors for WISC-III Verbal Scale and the CogAT6 Verbal Battery correlated r = *.87; whereas latent factors for WISC-III Performance Scale and the CogAT6 Nonverbal Battery correlated only* r = *.64. However, latent general factors on the two batteries correlated at least* r = *.97. It is argued that this correlation is plausible. Implications of these results for the interpretations of scores on both tests are discussed.*

The Wechsler Intelligence Scales are the most widely used intelligence tests in the world (Kaufman, 2000). The six editions of the Cognitive Abilities Test (CogAT) have been used in schools throughout the United States, Canada, and western Europe since the 1970s (for overviews of the fifth edition of CogAT, see Anastasi & Urbina, 1997; Linn & Gronlund, 2000). However, there are no published reports of the concurrent validity of the individually administered Wechsler scales and the group-administered CogAT. The purpose of this study was to investigate the relationships between recent editions of each test: Form 6 of the Cognitive Abilities Test (CogAT6; Lohman & Hagen, 2001a) and the Wechsler Intelligence Scale for Children—Third Edition (WISC-III; Wechsler, 1991).

Method

Sample

It is expensive and time consuming to administer individual intelligence tests. Consequently, studies that report correlations between individually administered tests and other variables often must rely on records of students who were referred for individual testing. Although reasonably large samples can be obtained by using these student records, such samples are necessarily unrepresentative of the larger population of school children. The ideal, of course, is obtain the sort of representative samples that test publishers secure at considerable cost when norming a test. Constraints of budget, parental/student consent, and school cooperation made it impossible for most investigators to obtain this type of sample.

The goal for this study was to obtain a sample of students that would be somewhere between these two extremes. Costs precluded testing more than 100 students. This sample size, although sufficient for estimating correlations, is uncomfortably small for confirmatory factor analyses. We sought to mitigate this concern by first establishing the extent to which the patterns of relationships among the subtest scores observed in our sample of volunteers differed from the norm samples for each test. We did this by following Bollen's (1989) recommendations (discussed later) for establishing the degree of equivalence between two covariance matrices.

We also sought to eliminate differences in scale scores that are introduced when students take different levels of a vertically equated test that is normed on different students in different grades (Kolen & Brennan, 1995). Equating error, sampling error, and systematic differences in content across levels of tests can attenuate correlations between two tests when scores from examinees who take different test levels are combined in one sample. We controlled this variation by testing only students in one grade at approximately the same time of year. Grade six was chosen because it was near the midpoint of the population on which the CogAT6 was normed and because many schools administer the CogAT at this grade.

[‡] David F. Lohman, College of Education.
I am grateful to Dr. Doug Becker and Ms. Lisa Morig of Riverside Publishing Company for coordinating data collection, and Dawn Bramer at the Belin-Blank Center for recruiting examiners and supervising their work. I also thank Patricia Martin for assistance in preparing the manuscript.

Volunteers were recruited from two middle schools in two industrial midwestern cities. Twenty-one of these students attended a public middle school in one city and 70 attended a Catholic middle school in another city. Both schools served neighborhoods that varied substantially in SES. Although information on parental income and education was not available to us, both principals characterized their students as predominantly middle class. Eighty-two students identified themselves as White, five as multiracial, two as Black, and two as Asian American.[1] The final sample consisted of 49 female and 42 male sixth-grade students. Average age was 11 years 10 months.

Instruments

WISC-III. The WISC-III is a revision of a series of tests that was first published as the Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949) and later revised as the WISC-R (Wechsler, 1974). Although subtests were selected for inclusion in the battery on the basis of previous empirical and clinical work, Wechsler's test batteries generally do not conform well to any one theory of intelligence (Kaufman, 2000).

The WISC-III contains ten standard and three supplemental subtests. These 13 tests are hypothesized to measure four factors: Verbal Comprehension (subtests Information, Similarities, Vocabulary, and Comprehension), Perceptual Organization (subtests Picture Completion, Picture Arrangement, Block Design, Object Assembly, and supplemental subtest Mazes), Freedom from Distractibility (subtest Arithmetic and supplemental subtest Digit Span), and Processing Speed (subtest Coding and supplemental subtest Symbol Search). Hierarchical confirmatory factor analyses confirm that these four factors can be identified from ages 6 to 16 (Keith & Witta, 1997).

In the study reported here, only the 10 standard WISC-III subtests were administered to all examinees. Therefore, the smaller Freedom from Distractibility and Processing Speed factors could not be identified. When supplemental subtests are not administered, the 10 standard subtests are typically combined into a Verbal Scale (Information, Similarities, Arithmetic, Vocabulary, and Comprehension) and a Performance Scale (Picture Completion, Coding, Picture Arrangement, Block Design, and Object Assembly) (Wechsler, 1991).

Internal consistency estimates of reliability for WISC-III subtests range from .68 (Object Assembly) to .89 (Vocabulary) for 12-year-olds (Wechsler, 1991). The corresponding estimates for composite scores are .95 for the Verbal Scale, .91 for the Performance Scale, and .96 for the Full Scale score. Parallel forms reliability coefficients are given by correlations between the WISC-III and the WISC-R for a sample of 206 children who were administered both tests. These range from .42 (Picture Arrangement) to .80 (Information). Correlations between the composite scores on the WISC-III and the WISC-R were .90 for the Verbal Scale, .81 for the Performance, and .89 for the Full Scale score (Wechsler, 1991).

CogAT6. Form 6 of the CogAT is the latest revision of a test that was first published in 1954 as the Lorge-Thorndike Intelligence test (Lorge & Thorndike, 1954). The first edition of the CogAT Primary Battery was published in 1968. The first edition of the complete test was published in 1971. Like all subsequent editions, it had a shorter, orally administered Primary Battery for grades K-3 and a longer Multilevel Battery suitable for grades 3-12. The test was jointly standardized with the Iowa Tests of Basic Skills (ITBS) at grades K-9 and with the Tests of Achievement and Proficiency (TAP).

Like the Wechsler scales, the CogAT has undergone several important revisions over the years. Also like the Wechsler scales, early editions of the test relied more on the author's definition of the construct the test aimed to measure than a particular model of human abilities. The stated goal was to measure "the ability to use and manipulate abstract and symbolic relationships" in each of three symbol systems: verbal, quantitative, and figural/spatial (Thorndike & Hagen, 1971, p. 3). Hierarchical factor analyses of the ten subtests in Form 1 Multilevel Battery showed a large "abstract reasoning" factor and three content-based group factors. However, the quantitative reasoning factor, which was clearly present at the group factor level, was only weakly distinguished from the general reasoning factor in the hierarchical factor analyses.

Recent editions of the CogAT have made more explicit this correspondence between the internal structure of the test and hierarchical models of abilities, particularly the Cattell-Horn-Carroll (CHC) three-stratum theory of cognitive abilities. CHC theory combines Cattell-Horn's *Gf-Gc* (Cattell, 1971; Horn, 1989) and Carroll's three-stratum theories of human abilities (Carroll, 1993). Modifying slightly the construct statement of Form 1, the authors of Form 6 claim that the CogAT focuses on the *g* factor at the third stratum of the CHC theory and the stratum II fluid

[1] The relative paucity of minority students (approximately 10 percent) is not necessarily bad when the goal is to estimate factor structures rather than means. If covariance structures are the same within each group (as, for example, is commonly the case when comparing Whites and Blacks), then the relative composition of the sample makes no difference. If, on the other hand, covariance structures differ across ethnic groups, then a more ethnically diverse sample is better only if sample size is large enough to permit within-group factor analyses.

reasoning (Gf) abilities that load most highly on *g* (Lohman & Hagen, 2001a). Carroll (1993) argues that the Gf factor is defined by three reasoning abilities: (1) *sequential reasoning*—verbal, logical, or deductive reasoning; (2) *quantitative reasoning*—inductive or deductive reasoning with quantitative concepts; and (3) *inductive reasoning*—typically measured with figural tasks. These correspond roughly with the three CogAT6 batteries: verbal reasoning, quantitative reasoning, and figural/nonverbal reasoning. Each of these three reasoning abilities is estimated from two tests in the Primary Battery (grades K-2) and from three tests in the Multilevel Battery (grades 3-12).

K-R 20 reliabilities of the Verbal, Quantitative, and Nonverbal scores range from .86 to .92 for the Primary Battery and from .94 to .95 for the Multilevel Battery (Lohman & Hagen, 2002). A general ability score is estimated from the average of scale scores across the three batteries. K-R 20 reliabilities of this composite score average .96 for the Primary Battery and .98 for the multilevel battery. Correlations between composite scores on two forms of the test administered two weeks apart average .92.

Three types of evidence on the validity of CogAT6 are presented in the *Research Handbook* (Lohman & Hagen, 2002). The first type of evidence is based on the content of test tasks, particularly on the cognitive processes examinees typically use to solve them. Cognitive psychologists have studied all of the tasks used in the test, some extensively so (Lohman, 2000). A framework based on Sternberg's (1986) theory of reasoning was used to guide the construction of test items so as to enhance construct-relevant variance and reduce construct-irrelevant variance. This theoretical framework was also used to inform interpretations of psychological constructs measured by the test, particularly their implications for instruction.

The second type of evidence presented in the CogAT6 *Research Handbook* is based on the internal structure of the test. A series of confirmatory factor analyses were performed on the correlation matrices for each of the three levels of the primary battery and the eight levels of the multilevel battery. Verbal, quantitative, and figural/nonverbal reasoning factors were extracted at level 1. A general factor was then extracted at the second level, and the two factor-pattern matrices were combined into a single hierarchical factor matrix using the procedures of Schmid and Leiman (1957). The median standardized root mean square residual for these models was .0143. For the Primary Battery (grades K-2), the quantitative factor identified at level 1 was completely subsumed by the general factor in the hierarchical model. For the Multilevel Battery, the loadings for the residualized quantitative factor were small but significantly greater than zero at all levels, especially in grades 7-12. Although the pattern of factor loadings was consistent across test levels with each battery, multiple-group confirmatory factor analyses were not performed.

The third type of validity evidence is based on the relationship between performance on CogAT and on other tests, particularly measures of school achievement. CogAT6 was co-normed with the Iowa Tests of Basic Skills on 149,798 students in grades K-8 and with the Iowa Tests of Educational Development on 30,740 students in grades 9-12. The concurrent prediction of achievement was quite high. Average correlations with the ITBS Composite score were .83 for the Verbal Battery, .78 for the Quantitative Battery, .71 for the Nonverbal Battery, and .86 for the CogAT6 Composite. Differential validity was also observed for the CogAT6 Verbal score (which showed highest correlations with reading and language tests on the achievement batteries) and for the CogAT6 Quantitative score (which showed highest correlations with mathematics test on the achievement batteries). Other studies reported in the *Research Handbook* show substantial prediction of achievement over time. For example, CogAT5 scores obtained in grade 4 predicted achievement test scores in grades 6 and 9 about as well as grade 4 achievement test scores predicted future achievement (Lohman & Hagen, 2002). Scores from earlier editions of the test also have been correlated with the Stanford-Binet Intelligence Scale and other group-administered ability tests such as the Differential Aptitude Tests (see Thorndike & Hagen, 1974, 1987; also footnote 3). However, there are no published reports of the concurrent validity of CogAT5 or CogAT6 with an individually administered intelligence test.

Procedure

Middle schools within a 75-mile radius of the five-member examiner team were identified. One was a public middle school that planned to administer CogAT6 to the 244 students in the sixth grade in November of 2001. After gaining approval of the district and school administrators, recruitment fliers were distributed to all students. Participants were offered a coupon that could be redeemed for a pizza. Parents were offered information on their child's learning abilities and study strategies that might assist them. Only 21 students agreed to participate.

The recruitment strategy for the second school thus took a different form. A Catholic middle school in a different city was identified that enrolled 139 sixth-grade students. The school was offered free CogAT6 testing with the proviso that they would try to achieve a participation rate of at least 50% of their sixth graders. The need for a representative sample was emphasized. In all, 70 students were tested in November and December of 2002.

Parent/guardian and student consent was obtained for each student in both schools. The ten subtests in the WISC-III standard battery were administered by one of five trained examiners using Wechsler's (1991) procedures to score subtests and compute Verbal, Performance, and Full Scale summary scores. Protocols were double-checked by a colleague who teaches courses in individual intelligence testing. Scoring was double-checked by a clerk. Within approximately 2-3 weeks of taking the WISC-III, students were administered level D of CogAT6 by their regular classroom teachers.

Statistical Analyses

The goals of the analyses were more modest than in most applications of confirmatory factor analysis (CFA). Our purpose was not to determine the most defensible factor structure for either the WISC-III or CogAT6. Rather, we sought to investigate relationships between these two test batteries in a small and not-too-unrepresentative sample of 91 student volunteers. The first set of analyses attempted to establish the representativeness of our sample. If the patterns of relationships among WISC-III subtests or among CogAT6 subtests observed in our sample differed from those reported by the test authors, then relationships between scores on the two test batteries are unlikely to generalize. If, on the other hand, we found congruence between the patterns of relationships among subtest scores in our sample and those reported by the two test publishers, then relationships we observe between the two test batteries have greater plausibility.

Univariate means and standard deviations provide some evidence on the on the extent to which the study sample faithfully represents the population. The most important concern for a correlational study, however, is whether the patterns of relationships among subtests in each battery mirror population covariances. The best estimate of the population covariance matrix for the tests in each battery is given in the standardization sample. For the WISC-III, we estimated the covariance matrix from the correlation matrix and subtest standard deviations for 12-year-olds (Wechsler, 1991). For the CogAT, we estimated the covariance matrix from the correlation matrix and subtest standard deviations for Level D (Lohman & Hagen, 2002).

We then compared the covariance matrices for our sample with the corresponding matrix reported by the test authors. Bollen (1989) discusses methods for performing this task. The most stringent procedure is simply to test the equality of the two covariance matrices. For even moderately large matrices, however, the probability that all k variances and $k(k - 1)/2$ covariances will be the same is exceedingly unlikely (Bollen, 1989). At the other extreme, one can test whether the factor structures of the two matrices have the same general form. Two models are said to have the same form if the model for each group has the same parameter matrices with the same dimensions and the same location of fixed, free, and constrained parameters. For example, two confirmatory factor models would have the same form if both specified three factors and the same paths between factors and observed variables. However, factor loadings and correlations among the factors could vary. Different degrees of equivalence may be identified between these two extremes (Bollen, 1989).

Because both the WISC-III and the CogAT6 have established factor structures, we first attempted to fit these hypothesized factor structures to covariance matrices extracted from the standardization tables. We were thus able to test whether the expected factor pattern described both the population covariance matrix and the corresponding covariance matrix for our sample. In both cases, we started with a fairly stringent definition of equivalence—same form, same factor loadings, and same factor intercorrelations—and then relaxed these restrictions as necessary.

For the CogAT6 data, we assumed that the 9 variances and 36 covariances in the standardization covariance matrix could be treated as population values. This assumption seemed reasonable, given the size and representativeness of the standardization sample (N = 13,407 sixth graders). We then tested whether the parameters of the model fit to these data could also describe the CogAT6 data in our sample.

The procedure used with the WISC-III data was a bit different. The assumption that the 45 covariances and 10 variances in the WISC-III standardization matrix could be treated as population values seemed unlikely, even for a representative sample of 200 twelve-year-olds. Therefore, we tested whether a common model could be fit to the 200-subject standardization data set and the 91-subject sample data set.

Once the degree of equivalence between sample and standardization covariance matrices was established for each test battery, we proceeded to estimate relationships between latent variables on the WISC-III and CogAT6 in models that combined the two test batteries. All confirmatory factor models were tested using AMOS 4.0 (Arbuckle, 1999). Factors were extracted from covariance matrices using maximum likelihood. Model fit was evaluated using the two-index strategy recommended by Hu and Bentler (1999). Root mean square error of approximation (RMSEA) was used as the absolute fit index. As recommended by Steiger and Lind (1980), upper and lower bounds of RMSEA were also reported. The Tucker-Lewis Index (TLI; also called the Non-normed Fit Index) was used as the incremental fit index. Hu and Bentler (1999) recommend that, when using

Table 1
Means, Standard Deviations, and Correlations (N=91)

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
CogAT 6																												
1 Verbal Classification	16.3	3.08	1.00	.61	.63	.50	.26	.39	.56	.37	.56	.77	.46	.53	.65	.59	.55	.43	.56	.35	.42	.24	.23	.42	.28	.63	.46	.65
2 Sentence Completion	16.0	3.32	.61	1.00	.66	.51	.39	.40	.53	.45	.50	.77	.50	.56	.68	.54	.53	.47	.57	.33	.18	.28	.15	.34	.23	.62	.35	.58
3 Verbal Analogies	19.0	3.99	.63	.66	1.00	.57	.44	.42	.53	.40	.49	.87	.54	.52	.72	.60	.61	.44	.60	.37	.38	.20	.22	.41	.24	.66	.43	.65
4 Quantitative Relations	19.7	4.31	.50	.51	.57	1.00	.62	.61	.47	.49	.58	.51	.83	.55	.71	.49	.30	.56	.39	.26	.13	.24	.12	.45	.12	.52	.32	.50
5 Number Series	15.6	3.40	.26	.39	.44	.62	1.00	.57	.44	.52	.63	.42	.82	.60	.71	.39	.15	.57	.44	.27	.21	.26	.13	.48	.14	.49	.36	.50
6 Equation Building	11.8	2.96	.39	.40	.42	.61	.57	1.00	.36	.48	.55	.43	.77	.52	.65	.37	.16	.49	.34	.19	.01	.41	.05	.39	.11	.41	.30	.42
7 Figure Classification	20.1	4.97	.56	.53	.53	.47	.44	.36	1.00	.70	.68	.58	.50	.82	.71	.40	.30	.36	.40	.33	.32	.21	.26	.40	.27	.46	.43	.53
8 Figure Analogies	19.3	4.48	.37	.45	.40	.49	.52	.48	.70	1.00	.65	.48	.58	.86	.73	.40	.26	.34	.38	.22	.32	.22	.20	.45	.30	.41	.44	.51
9 Figure Analysis	10.8	3.47	.56	.50	.49	.58	.63	.55	.68	.65	1.00	.59	.66	.83	.79	.47	.25	.51	.53	.33	.31	.15	.21	.57	.26	.55	.45	.59
10 Verbal SAS	109.7	12.54	.77	.77	.87	.51	.42	.43	.58	.48	.59	1.00	.56	.67	.83	.71	.68	.53	.76	.41	.45	.29	.26	.49	.37	.78	.55	.79
11 Quantitative SAS	110.9	13.87	.46	.50	.54	.83	.82	.77	.50	.58	.66	.56	1.00	.69	.87	.51	.28	.68	.49	.25	.14	.39	.15	.54	.19	.58	.42	.59
12 Nonverbal SAS	110.4	15.31	.53	.56	.52	.55	.60	.52	.82	.86	.83	.67	.69	1.00	.91	.52	.38	.50	.56	.35	.40	.25	.27	.60	.41	.59	.57	.69
13 Composite SAS	111.5	14.12	.65	.68	.72	.71	.71	.65	.71	.73	.79	.83	.87	.91	1.00	.65	.50	.65	.68	.38	.37	.36	.26	.63	.37	.74	.59	.79
WISC III																												
14 Information	12.5	3.04	.59	.54	.60	.49	.39	.37	.40	.40	.47	.71	.51	.52	.65	1.00	.56	.53	.73	.44	.44	.05	.06	.33	.16	.83	.30	.68
15 Similarities	12.4	2.39	.55	.53	.61	.30	.15	.16	.30	.26	.25	.68	.28	.38	.50	.56	1.00	.41	.60	.40	.46	.04	.17	.31	.30	.73	.37	.66
16 Arithmetic	11.9	3.52	.43	.47	.44	.56	.57	.49	.36	.34	.51	.53	.68	.50	.65	.53	.41	1.00	.63	.28	.20	.20	.14	.43	.16	.75	.35	.65
17 Vocabulary	11.4	2.74	.56	.57	.60	.39	.44	.34	.40	.38	.53	.76	.49	.56	.68	.73	.60	.63	1.00	.47	.42	.17	.17	.46	.28	.87	.44	.78
18 Comprehension	12.5	3.57	.35	.33	.37	.26	.27	.19	.33	.22	.33	.41	.25	.35	.38	.44	.40	.28	.47	1.00	.25	.11	.14	.23	.15	.69	.26	.57
19 Picture Completion	11.5	2.92	.42	.18	.38	.13	.21	.01	.32	.32	.31	.45	.14	.40	.37	.44	.46	.20	.42	.25	1.00	.00	.24	.36	.41	.44	.57	.59
20 Coding	11.4	3.32	.24	.28	.20	.24	.26	.41	.21	.22	.15	.29	.39	.25	.36	.05	.04	.20	.17	.11	.00	1.00	.20	.31	.23	.16	.53	.39
21 Picture Arrangement	36.8	9.73	.23	.15	.22	.12	.13	.05	.26	.20	.21	.26	.15	.27	.26	.06	.17	.14	.17	.14	.24	.20	1.00	.39	.41	.17	.67	.49
22 Block Design	12.4	3.51	.42	.34	.41	.45	.48	.39	.40	.45	.57	.49	.54	.60	.63	.33	.31	.43	.46	.23	.36	.31	.39	1.00	.62	.46	.80	.73
23 Object Assembly	10.2	2.85	.28	.23	.24	.12	.14	.11	.27	.30	.26	.37	.19	.41	.37	.16	.30	.16	.28	.15	.41	.23	.41	.62	1.00	.26	.78	.60
24 Verbal IQ	112.7	13.81	.63	.62	.66	.52	.49	.41	.46	.41	.55	.78	.58	.59	.74	.83	.73	.75	.87	.69	.44	.16	.17	.46	.26	1.00	.44	.86
25 Performance IQ	109.4	14.42	.46	.35	.43	.32	.36	.30	.43	.44	.45	.55	.42	.57	.59	.30	.37	.35	.44	.26	.57	.53	.67	.80	.78	.44	1.00	.83
26 Full Scale IQ	112.0	13.12	.65	.58	.65	.50	.50	.42	.53	.51	.59	.79	.59	.69	.79	.68	.66	.65	.78	.57	.59	.39	.49	.73	.60	.86	.83	1.00

Note. CogAT6 = Level D of the *Cognitive Abilities Test* (Form 6); WISC-III= *Wechsler Intelligence Scale for Children –Third Edition.*

maximum likelihood, a cutoff value close to .95 for TLI and close to .06 for RMSEA are needed before one can conclude that there is a relatively good fit between the hypothesized model and the observed data.

Comparison of nested models was based primarily on a comparison of χ^2s for the two models. If the change in χ^2 was statistically significant, then the model with the smaller χ^2 (and smaller *df*) was retained. If the change in χ^2 was not statistically significant, then the more parsimonious model (the model with the larger *df*) was chosen.

Results

Descriptive statistics

Complete data on the nine CogAT6 subtests and the ten WISC-III subtests were obtained for all 91 students. Univariate distributions for each subtest and summary score were plotted and then examined for outliers. None were identified. Average skewness across the 19 subtests was -0.52 (range -1.69 to .46). Average kurtosis was .45 (range -.87 to 4.42). Kurtosis exceeded 2.0 for the CogAT6 Verbal Classification and Sentence Completion subtests. We also checked for multivariate outliers by computing the Mahalanobis distance between each vector of 19 scores (ten WISC-III subtests plus nine CogAT6 subtests) and the centroid. None exceeded the criteria of $p < .001$ suggested by Tabachnick and Fidell (1996). Multivariate normality was assessed with Mardia's (1970) coefficient of multivariate kurtosis. The value of 21.72 had a critical ratio of 3.66, which is a somewhat high. A rank-order transformation on the two CogAT6 subtests with the highest univariate kurtosis estimates reduced Mardia's coefficient to 12.39 (critical ratio 2.09). Subsequent models were tested with and without transforming the two CogAT6 subtests. Differences were generally in the third decimal place, and so the results reported here are based on the untransformed CogAT6 raw scores and the WISC-III scale scores.

Univariate means and standard deviations are reported in the first two columns of Table 1. The students in this sample were above average in ability. Mean CogAT6 SAS score was 111.5; mean WISC-III Full Scale IQ was 112. Average scores that depart from the population mean can signal restricted score variability. Although there was some restriction in range, variability of scores was quite good for both batteries. Standard deviations ranged from 12.5 to 15.5 across the four CogAT6 SAS scores and from 13.1 to 14.4 on the three WISC-III IQ scores. (Population *SD*s are 15 for the WISC-III and 16 for the CogAT6 batteries.)

Modeling the WISC-III Data

The first step in our analysis was to fit the hypothesized two-factor model to the covariance matrix for 200 twelve-year-olds reported in the WISC-III manual. We then attempted to fit the same model to our data and then to both matrices simultaneously.

The general form of the two-factor model for the WISC-III is shown in Figure 1. (Note, however, that the factor loadings in Figure 1 are for the last model in this series.) As expected, this model (Model I-a in Table 2) fit the standardization covariance matrix quite well ($\chi^2 = 54.1$, $df = 34$, Tucker-Lewis Index = .973, RMSEA = .054). Model I fit the covariance matrix for the sixth graders in our sample only slightly less well ($\chi^2 = 45.9$, $df = 34$, Tucker-Lewis Index = .947, RMSEA = .062). This establishes what Bollen (1989) calls *form equivalence*. It says that the same basic model can be fit to both data sets. However, the parameters of the models are allowed to vary across the two samples.

More stringent definitions of equivalence require that some or all of the model parameters be the same (Bollen, 1989). We tested this in a series of a simultaneous multiple group analyses on the two covariance matrices. The baseline for these models was the sum of the χ^2 and *df* for the analyses in which Model I was separately fit to each matrix. As shown in Table 2, the baseline model (Model I-c) has a χ^2 of 100.1 with 68 *df*.

Table 2
Confirmatory factor models for the Wechsler Intelligence Scale for Children—Third Edition

Model and Sample	χ^2	*df*	TLI	RMSEA	
				Value	Range
Independent Models					
I. Factor loadings and V-P covariance free					
I-a. Standardization ($N = 200$)	54.1	34	.973	.054	.024-.081
I-b. Grade 6 ($N = 91$)	45.9	34	.947	.062	.000-.105
Simultaneous Models (Grade 6 + Standardization)					
I-c. Factor loadings and V-P covariance free	100.1	68	.967	.040	.022-.057
I-d. Factor loadings and V-P covariance constrained	118.7	77	.962	.043	.027-.058
I-e. Factor loadings constrained, V-P covariance free	113.3	76	.966	.041	.024-.056

Note. TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation.

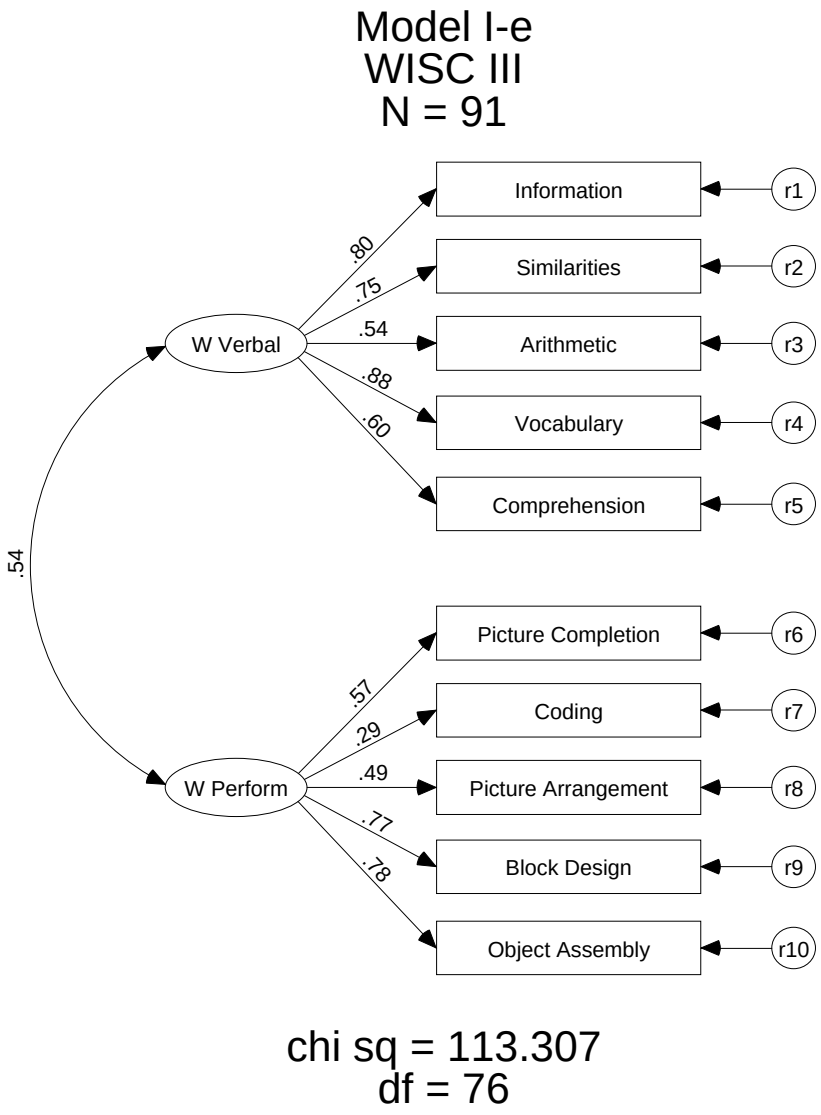

Figure 1. Model I-e. Final model for the WISC-III. Factor loadings were constrained to be equal across groups, but the correlation between the two factors was allowed to vary. Standardized loadings shown are for the sample of 91 students.

We began by testing a strong equivalence model. In this model, the ten unstandardized factor loadings and the covariance between the two factors were constrained to be the same in the models fit to both matrices. This is called Model I-d in Table 2. The χ^2 for this model was 118.7, an increment of 18.6 (9 *df*) over the sum of the χ^2s for the analysis in which Model I was separately fit to each data set. This increment in chi square is significant at $p = .05$ level. Therefore, the strong equivalence model does not hold.

The hypothesis that the covariance between the Verbal and Performance factors is the same in both samples seems unlikely, given the moderate restriction in range for the test sample. This would tend to lower the correlation between the two factors. Therefore, we relaxed the restriction of equal covariance between the factors. The revised model (Model I-e in Table 2) gave a χ^2 of 113.3. The increase in χ^2 over the baseline χ^2 of 100.1 was not significant ($\Delta\chi^2 = 13.2$, $df = 8$, $p > .05$). Factor loadings for this model are shown in Figure 1. We conclude that, with the minor modification of allowing the Verbal and Performance factors to correlate differently in the two samples, the same model fit the covariance matrix for our sixth-grade sample and the standardization sample of 12-year-olds.

Modeling the CogAT6 Data

Slightly different procedures were used to model the CogAT6 data. Instead of fitting a common model to both data sets, we assumed that the covariances among tests in the standardization sample could be treated as population values. A three-factor model was then fit to

Table 3
Confirmatory factor models for Level D of the Cognitive Abilities Test (Form 6)

Model and Sample	χ^2	*df*	TLI	RMSEA Value	RMSEA Range
Baseline Models					
II. Factor loadings and V, Q, N covariances free					
II-a. Standardization sample (*N* = 13,407)	834.4	24	.985	.050	.047-.053
II-b. Grade 6 (*N* = 91)	44.7	24	.930	.098	.051-.142
II-c. Grade 6 (*N* = 91) 1 pair of residual variances covary[a]	35.6	23	.956	.078	.010-.126
Constrained Models (Grade 6 sample only)					
II-d. Factor loadings and covariances constrained; 1 pair residual variances covary [a]	66.0	32	.914	.109	.071-.146
II-e. Factor loadings constrained, factor covariances free; 1 pair residual covariances covary [a]	48.1	29	.946	.086	.039-.127

Note. TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation.

[a]Residual variances for CogAT6 Verbal Classification and Number Series were allowed to covary.

these data. We then tested whether a model with the same factor loadings and factor covariances would also fit the sample data. The baseline for these comparisons was the same three-factor model for the sample data in which factor loadings and factor covariances were allowed to vary freely.

Figure 2 shows the hypothesized three-factor structure for the CogAT6 Multilevel battery. (Once again, the factor loadings in Figure 2 are for the last model in the series.) The fit statistics for this series of models are shown in Table 3. As expected, the hypothesized factor structure (called Model II-a) fit the standardization covariance matrix quite well (Tucker-Lewis Index = .985; RMSEA = .050). The same model was then fit to the covariance matrix for the sample of 91 sixth graders. Factor loadings and covariances were allowed to vary freely. Model II-b fit the sample of 91 sixth graders reasonably well (Tucker-Lewis Index = .930; RMSEA = .098). Fit improved when we allowed the residuals for Verbal Classification and Number Series subtests to covary.[2] This is called Model II-c in Table 3.

Model II-d tests the hypothesis that factor loadings and correlations can be constrained to their population values. All nine unstandardized regression weights and the three factor covariances were fixed to the values obtained in fitting Model II to the standardization data. As shown in Table 3, the χ^2 for fitting Model II-d to the sample data was 66.0, which represents a significant increase in χ^2 over the baseline χ^2 of 35.6. Thus, the assumption of strong equivalence must be rejected.

Once again, we relaxed the restriction that covariances among the factors be the same in the two models. This is called Model II-e in Table 3. Factor loadings for the model are shown in Figure 2. This resulted in a large reduction in χ^2 from Model II-d. As shown in Table 3, the resulting χ^2 of 48.1 (29 *df*) represents an increase in of χ^2 of 12.5 over the baseline Model II. With 6 *df*, this increment in χ^2 is not significant. Therefore, we conclude that the model that fits our CogAT6 data is not markedly different from the model that fits the standardization data.

Interbattery Analyses

We have now established that the relationships among subtests on each test battery that we observed in our sample sixth graders are reasonably congruent with those observed in the larger and more representative standardization samples for the two test batteries.

Two different types of interbattery analyses are reported. First, we report correlations between the three WISC-III summary scores (Verbal, Performance, and Full Scale scores) and the four CogAT6 summary scores (Verbal, Quantitative, Nonverbal, and Composite Standard Age scores). Second, we report correlations between latent factors corresponding to each of these scores that were represented in our CFA models for each battery.

WISC-III Scale score, CogAT6 SAS correlations. Users of either the WISC-III or the CogAT6 typically report scale scores on the WISC-III or standard age scores (SAS) on the CogAT6. Thus, the most direct comparison of the two batteries is given by the correlations between the WISC-III Verbal, Performance, and Full Scale scores and the four CogAT6 SAS scores. These correlations are reported in the last three rows of Table 1.

The first question to be addressed is the extent to which the general ability scores on the two test batteries are correlated. General ability is estimated by the Full Scale score on the WISC-III and by the

[2] Given the sample size, we worried about over-fitting the data. A reviewer, on the other hand, worried about model instability given the RMSEA of .098. In the end, however, it made little difference in the relationships between latent variables on the two tests.

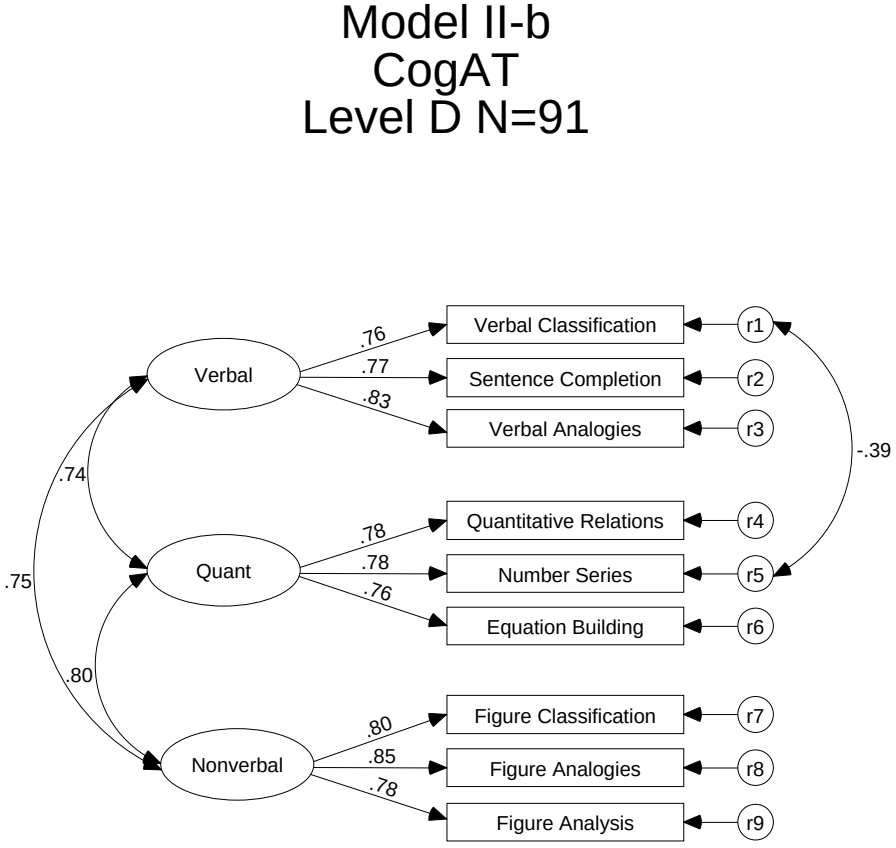

Figure 2. Model II-b. Final model for the CogAT6. Factor loadings were constrained by setting the unstandardized paths to the values obtained in fitting Model II to the standardization covariance matrix. Covariances between factors were allowed to vary. Standardized loadings and factor correlations shown are for the sample of 91 students.

Composite SAS score on CogAT. The correlation of r = .79 compares favorably with the correlation of r = .76 Thorndike and Hagen (1974) reported between SAS scores for the first edition of CogAT and the Stanford-Binet, Form L-M.[3]

The patterns of correlations between the CogAT6 Verbal, Quantitative, and Nonverbal batteries and the WISC-III Verbal and Performance scores were examined next. CogAT6 Verbal correlated highest with WISC-III Verbal (r = .78). CogAT6 Nonverbal SAS correlated about equally with both WISC-III Verbal (r = .59) and WISC-III Performance (r = .57) scale scores. The CogAT6 Quantitative SAS also correlated r = .58 with the WISC-III Verbal but only r = .42 with the WISC Performance scale score.

Correlations between observed scores address the practical matter of how much discrepancy test users are likely to see in scores on the two tests. However,

[3] Since the first edition of CogAT did not report a composite score, we computed this correlation from the reported correlations between the Stanford-Binet and the CogAT Verbal, Quantitative, and Nonverbal scores of r = .75, .68, and .65, respectively, and from the average correlations among the three CogAT scores for the test levels represented in the study.

Table 4
*Correlations among Latent Variables (*Upper Diagonal*) and Observed Scores (*Lower Diagonal*) (N=91)*

		1	2	3	4	5
CogAT						
1	Verbal	—	**.72**	**.75**	**.87**	**.56**
2	Quantitative	.56	—	**.80**	**.63**	**.54**
3	Nonverbal	.67	.69	—	**.63**	**.64**
WISC						
4	Verbal	.78	.58	.59	—	**.55**
5	Performance	.55	.42	.57	.44	—

Note. Correlations among latent variables (in bold) above the diagonal and among observed variables below the diagonal. Observed scores are Standard Age Scores on CogAT6 and Scale Scores on the WISC-III. Latent variable correlations are based on Model II-b.

Table 5
Confirmatory factor models for the combined WISC-III–CogAT6 analyses

				RMSEA	
Model and Sample	χ^2	*df*	TLI	Value	Range
III-a. Constrain g-level 1 paths, 1 pair of CogAT6 residuals allowed to covary.[a]	257.7	148	.856	.981	.072-.109
III-b. Constrain g-level 1 paths, 4 additional between-battery pairs of residuals allowed to covary.[b]	213.5	144	.906	.073	.051-.093
III-c. Unconstrain g-level 1 paths, 4 additional between-battery pairs of residuals allowed to covary.[b]	187.8	141	.936	.061	.034-.082

Note. TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation.

[a]See Model II-c.

[b]The pairs were WISC-III Arithmetic residual with CogAT6 Quantitative residual, WISC-III Similarities residual with CogAT6 Verbal Analogies residual, WISC-III Coding residual with CogAT6 Equation Building residual, and WISC-III Block Design residual with CogAT6 Figure Analysis residual.

questions about relationships among the constructs measured by the two batteries are better addressed through confirmatory factor analysis models that estimate correlations between latent variables. These were estimated from a model like that shown in Figure 3, but without the two general factors. Instead, correlations among the five first-order factors were estimated.[4] The between-battery correlations are reported above the diagonal in Table 4. Observed score correlations among these variables from Table 1 are also reproduced below the diagonal for ease of comparison.

Correlations for the latent variables show that the two verbal batteries were most similar. Although the WISC-III Performance scale had its highest correlation with the CogAT6 Nonverbal, the other two CogAT6 batteries showed substantial correlations with the WISC-III Performance scale as well. Clearly, the WISC-III Performance scale cannot be equated with the CogAT6 Nonverbal Battery. Therefore, although there was a clear association between the two verbal batteries and a weaker association between the WISC-III Performance and CogAT6 Nonverbal batteries, the overriding factor seems to be the presence of a common general factor in both test batteries.

Confirmatory interbattery factor analysis. Since the major purpose of this study was to estimate the correlation between the latent general factors measured by the WISC-III and CogAT6, the final set of models are in some respects the most important. All of these models (see Figure 3) combine a hierarchical version of the basic CogAT6 model (left panel) with a hierarchical version of the basic WISC-III model (right panel). In the first model (Model III-a in Table 5), we set the variance of both general factors to 1.0 and constrained to be equal all paths from each *g* to its respective first order factors. This was done both to reduce the number of free parameters that needed to be estimated and to make the model conform to the way the general factor is actually estimated on both tests (i.e., as the average of scale scores across the separate batteries). The resulting correlation between the CogAT6 *g* and the WISC-III *g* was $r = 1.00$. However, as shown in the first row of Table 5, model fit statistics were poor (Tucker-Lewis Index = .856; RMSEA = .091). This suggests that the model may be unstable and the parameters not identified.

[4] Correlations were estimated assuming Model II-b for the CogAT6 data. They were also estimated assuming Model II-c. The average difference in correlation was .01. As would be expected, the largest difference (.03) was between CogAT6 Verbal and Quantitative factors.

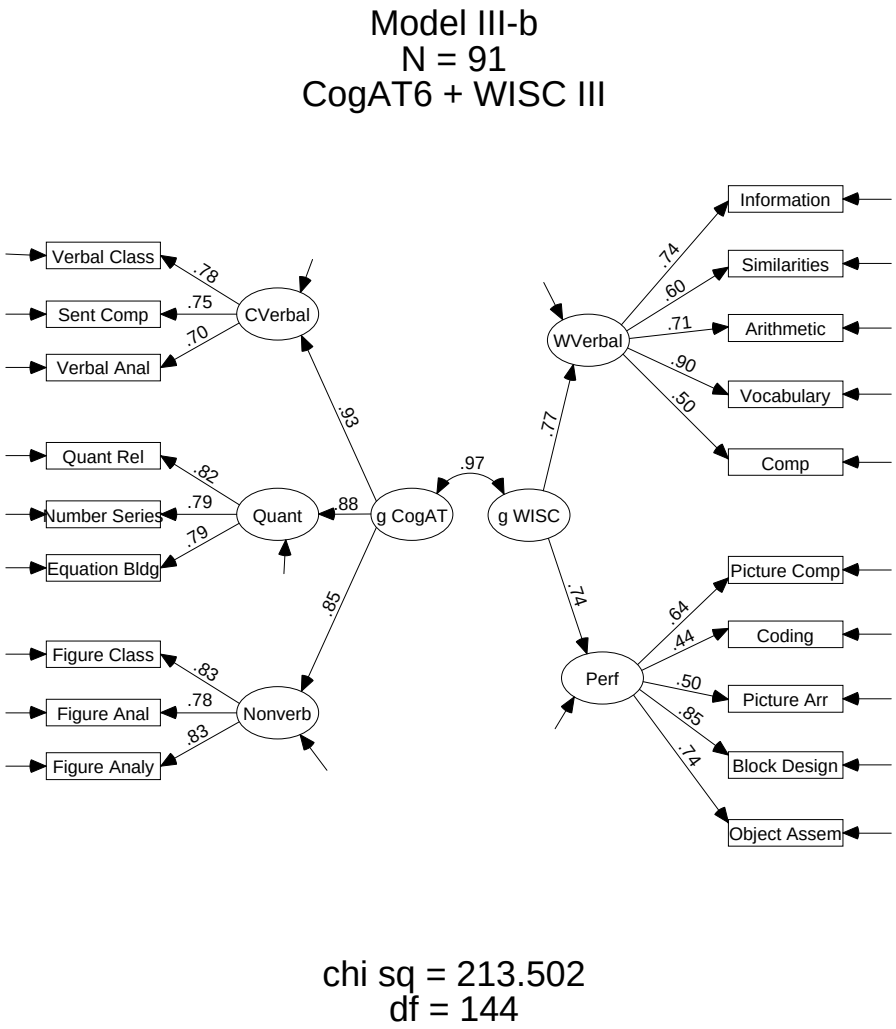

Figure 3. Model III-b. Confirmatory interbattery factor analysis. Standardized loadings and factor correlations shown are for the sample of 91 students. The unique and error variances of the measured variables are not shown to improve readability.

We checked the model in four ways. The first check involved relaxing restrictions on Model III-a to see if a better-fitting model gave a smaller correlation between the two general factors. Examination of the residual covariance matrix and the modification indices for Model III-a suggested that allowing four pairs of residuals to covary would improve fit. The pairs were WISC-III Arithmetic residual with CogAT6 Quantitative residual, WISC-III Similarities residual with CogAT6 Verbal Analogies residual, WISC-III Coding residual with CogAT6 Equation Building residual, and WISC-III Block Design residual with CogAT6 Figure Analysis residual. All four are psychologically plausible. This model (called Model III-b) is shown in Figure 3. The fit statistics for the Model III-b are shown in the second row of Table 5. Allowing the four residual covariances to be greater than zero improved the model fit somewhat. Further, as is shown in Figure 3, the correlation between the two *g* factors was reduced from $r = 1.00$ to $r = .97$.

Next, we also unconstrained the *g* to level 1 paths in both test batteries. This is called Model III-c in Table 5. As expected, fit statistics for this model were even better (Tucker-Lewis Index = .936; RMSEA = .061). However, the correlation between the two *g* factors remained $r = .97$. Further, the standardized regression path from *g* CogAT6 to CVerbal became 1.01 in this model. Although factor loadings greater than 1.0 sometimes are observed, here it seems to distort the meaning of the construct as well. Therefore, Model III-b (shown in Figure 3) seems preferable, even though it does not fit the data quite as well as Model III-c.

The second check involved specifying one *g* factor rather than two *g* factors in Models III-a to III-c. If the two *g* factors are almost perfectly correlated, then a

model with one *g* factor should fit as well as a model with two *g* factors. This was what we observed. Chi squares for models III-a, III-b, and III-c with one *g* factor were 257.7 (*df* = 149), 213.7 (*df* = 145), and 187.8 (*df* = 142), respectively. These χ^2s are virtually identical to the corresponding values in Table 5. The largest difference was 0.2.

The third check was an empirical test for model identification suggested by Jöreskog and Sörbom (1986) and recommended by Bollen (1989). The first step in this check is to analyze the covariance matrix and then save the predicted covariance matrix that is based on the estimates of the model parameters. The second step is to substitute the predicted covariances for the observed covariances and rerun the program. If the model is identified, then the new estimated model parameters should be identical to the original estimated model parameters. We performed this test on all three models (Model III-a, III-b, and III-c) and found that the model parameters were identical.

The fourth check did not attempt to confirm the integrity of the CFA models, but rather addressed the plausibility of a correlation of .97 to 1.00 between the two general factors. Latent factors on each test battery discard both error of measurement and test specificity. The disattenuated correlation between the WISC-III Full Scale score and the CogAT6 Composite score corrects for error of measurement in a score that only approximates *g*. Therefore, it sets a lower bound on the correlation between the two latent *g* factors. However, disattenuation requires choice of the most appropriate reliability coefficients. Parallel-forms reliability coefficients treat as error both form and occasion variance. Therefore, they are more appropriate than internal consistency estimates for estimating the relationship between constructs on different tests that were administered on different occasions. The reported parallel-forms reliabilities are .89 (WISC-R versus WISC-III for 12-year-olds)[5] and .92 (CogAT5 versus CogAT6 for Level D). Adjusting these values for the slight restriction of range in our sample gave reliabilities of .856 and .897 for the WISC-III Full Scale score and the CogAT6 Composite score, respectively. Using these reliabilities to correct the observed correlation of *r* = .79 gave a disattenuated correlation of *r* = .90 between WISC-III Full Scale score and CogAT6 Composite score. If the lower bound of the correlation between the latent *g* variables is .90, then an observed value of .97 is not implausible and a value of .99 or even 1.00 is not impossible.

[5] This correlation overestimates the reliability of the WISC-III since it is across all 13 subtests rather than the 10 used in this study.

Discussion

The purpose of this study was to investigate the concurrent validity of the Cognitive Abilities Test (Form 6) and the Wechsler Intelligence Scale for Children III. We first tested whether the pattern of relationships among subtests in each battery differed significantly from those reported in the standardization manuals for the two tests. The results showed only minor differences between the factor models that described our data and the corresponding standardization sets reported by the test authors.

Given this evidence on the representativeness of the within-test covariances in our data, we investigated relationships between scores on the two test batteries in three ways. First, we examined correlations between WISC-III scale scores and CogAT6 SAS scores. The WISC-III Full Scale score correlated *r* = .79 with the CogAT6 Composite score. The CogAT6 Verbal battery correlated highest with the WISC-III Verbal Scale Score, whereas the CogAT6 Nonverbal battery showed lower but equal correlations with the WISC-III Verbal and Performance scores. Correlations for the CogAT6 Quantitative battery were intermediate. Next, we examined correlations among the five latent variables that corresponded to the CogAT6 Verbal, Quantitative, and Nonverbal scores, and the WISC-III Verbal and Performance scores. These correlations mirrored the correlations between observed scores. The only noticeable difference was that the WISC-III Performance factor showed a slightly larger correlation with the CogAT6 Nonverbal factor than with the CogAT6 Verbal and Quantitative factors (correlations among observed scores did not show this differentiation). However, the WISC-III Performance scale clearly could not be equated with the CogAT Nonverbal Battery.

Finally, we investigated relationships between the general factors defined by the two test batteries in an interbattery factor analysis. This analysis showed that the general factors defined by the CogAT6 and WISC-III correlated at least *r* = .97. After several checks, we concluded that the general intellectual ability factor measured by the CogAT6 is the same general ability factor that is measured by the WISC-III. Even though the general factors appeared to be virtually coincident in this analysis, it is unlikely that they are exactly the same. In a larger sample, the correlation might well be less than the correlation of *r* = .97 that we observed in our best-fitting model. However, it would probably also be greater than *r* = .90. Nonetheless, the overlap is surely substantial.

This raises two questions: “Why?” and “So what?” Why might the correlation be so high? Unlike ability tests that were designed to measure primary abilities (in the Thurstone, 1938, tradition), or broad group

factors (as in the Horn-Cattell-Carroll theory [Carroll, 1993; Horn, 1989]), the CogAT was explicitly designed to measure abstract reasoning abilities. Although there is some debate about Gustafsson's (2002) claim that Inductive Reasoning = Gf = g, there no doubt that reasoning abilities are central to the definition of *g* (Carroll, 1993). Put differently, tests that aim to measure ability constructs that can be distinguished from *g* will in general have lower loadings on *g* than well-constructed tests explicitly designed to measure g.

Second, tests such as the CogAT that use multiple measures to estimate each ability score reduce the impact of item format on estimates of the abilities that they provide. Why does this matter? Every ability test measures something that generalizes to other tests that measure the same ability and something that is unique to the test (particularly the sample of items and their format) and the measurement occasion. Consider, for example, the factor loadings for the interbattery factor analysis that are shown in Figure 3. Verbal Classification, the first test in the CogAT6 battery, shows the largest loading on the CVerbal factor (.78). This means that the commonality for this test is $(.78)^2$ or .608. The uniqueness is 1 – .608 or .392—almost 40 percent of the total variance. For the WISC-III, the average loading of the five verbal tests on the WVerbal factor in Figure 3 is .69. This means that about 50 percent of the variance on each WISC-III verbal test is unique to the test. Adding more items to each test reduces error of measurement due to item sampling, but does not reduce task-specific variance due to item format. One can reduce the unwanted effects of task-specific factors by developing test items that maximize construct-relevant variance and by measuring the ability in as many different ways as possible. The CogAT authors report that item development was guided by research on the processes examinees use to solve different variants of each item type (Lohman & Hagen, 2002). This should have helped reduce the impact of construct-irrelevant variance in each test. Aggregating scores across three different tests reduces the impact of task-specific variance on each of the three reasoning abilities (Verbal, Quantitative, and Nonverbal/Figural). Aggregating scores a second time across the three reasoning abilities reduces the construct-specific variance in the composite score. Therefore, it is not at all implausible that the CogAT6 Composite score is a very good measure of g.

The "So what?" question has several answers. Only three will be mentioned here. If a group-administered test measures the same ability as an individually administered test, then it could be argued that the group test could be used to estimate this ability when it is impossible or impractical to test each child individually. Here, the concordance between the individual and group test is at the level of the general factor, and secondarily at the level of the two verbal factors. However, the construct measured by the WISC-III Performance scale is not the same as the construct measure by the CogAT6 Nonverbal Battery. Whether a WISC-III Quantitative or Freedom from Distractibility factor (defined by Arithmetic and Digit Span) would measure the same thing as the CogAT6 Quantitative Battery could not be tested. Therefore, even though there appears to be substantial overlap at the level of the general factor, the two test batteries are clearly not exchangeable. Furthermore, the goal of an individual examination is usually much broader than the summary score that is produced. The observations of a skilled examiner cannot be duplicated on a group-administered test.

Nevertheless, some problems that an examiner can observe can also be detected on a group-administered test. For example, group tests are often criticized because they cannot detect whether a child misunderstands the directions for a subtest, breaks the point on her pencil, or gets lost on the machine-readable answer sheet. However, this type of criticism may not apply to CogAT6. Such confusions lead to inconsistencies in the way a child responds to items on a test battery or to different subtests in that battery. These inconsistencies are captured in a personal error of measurement for each examinee (Lohman & Hagan, 2002). This error is then used to construct confidence intervals around each examinee's test scores. Examinees whose behavior is consistent with the expectations of the scaling model across items and across subtests in a battery will have narrow confidence intervals around their scores. Those who respond inconsistently (for whatever reason) will have much larger confidence intervals. These procedures are a standard part of the CogAT6 score reports (Lohman & Hagen, 2001a). Users are warned not to make decisions about such students until a more dependable estimate of ability is obtained, either through retesting or through administration of an individual test. In addition, various caution indices warn of other threats to validity, such as when students appear to have adopted an extremely slow but accurate response style. In short, although the group test can never replace the individual test, it need not be as poor of a substitute as it sometimes is. Indeed, individually administered tests might well incorporate some of the procedures for cautioning scores now being used on group-administered tests.

The second implication concerns consequential validity (Messick, 1989). Scores on intelligence tests such as the WISC-III are a critical part of an individual psychological assessment. Indeed, there is a vast literature on the clinical interpretations of such tests (e.g., Kaufman, 1994), virtually all of which is aimed

at the professional psychologist. However, scores on intelligence tests are often misinterpreted as measures of innate potential or capacity by teachers, students, and parents (Anastasi & Urbina, 1997). Without specific guidance, they often have no idea how to make use of the scores the test offers in order to improve instruction. This is unfortunate because research shows that reasoning abilities moderate the effectiveness of different instructional methods more commonly than do profiles of more specific abilities (Corno, Cronbach, Kupermintz, Lohman, Mandinach, Porteus, & Talbert, 2002). This means that efforts to adapt instruction to individual differences are much more likely to succeed if they are based on students' abilities to reason in a symbol system rather than to perform other types of cognitive operations on them.

Group ability tests, on the other hand, are designed to be administered and used by teachers. Support materials for these tests are written primarily for teachers and counselors, not professional psychologists. Furthermore, interpretations are geared to the typical child, not to the child who requires special interventions. These support materials can help teachers and counselors understand not only what ability tests measure, but how they can use ability test scores to improve instruction for a much broader range of children than the handful who are referred for individual assessment. This is particularly true for CogAT6, which provides separate manuals for teachers and administrators (Lohman & Hagen, 2001b, 2001c) and an extensive online service to assisting teachers in interpreting score profiles on the test (see www.cogat.com). Therefore, if the general factor on the two tests is the same, then empirically validated uses of one the general score on test can inform uses and interpretations of the general score on the other test.

The final implication concerns the validity of the WISC-III Performance scale score. Whatever this heterogeneous collection of tests measures, it is not the figural reasoning construct measured by the Nonverbal Battery on CogAT6. Block Design had its highest correlation with the CogAT Nonverbal score, and so the differences between the batteries were not simply due to method variance (i.e., performance tests on WISC-III versus paper-and pencil tests on CogAT6). Hopefully the next revision of the WISC will improve the clarity of the construct measured by the performance subtests. Separation and better representation of the factor Keith and Witta (1997) call *Quantitative Reasoning* would also seem worthwhile for educators, given the importance of quantitative reasoning as a potential marker for *g* and as an aptitude for mathematics and science learning.

Finally, although the methods used in this study to establish the representativeness of the sample covariance matrices should be useful in other validity studies, the study is clearly limited by the sample size and the above-average performance of the group. However, replication on an average ability group would not be expected to reduce the observed relationship between general factors on the two tests. If anything, the general factor is typically stronger in samples of less-able students than in samples of more-able students (Detterman & Daniel, 1989). Nevertheless, a larger sample size would allow one to estimate with greater certainty the whether correlation between the general factors on these two batteries is closer to the estimated lower bound of $r = .90$ or the upper bound of $r = 1.00$. The evidence offered here is that it would be closer to the latter than the former value.

References

Anastasi, A., & Urbina, S. (1997*). Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Arbuckle, J. L. (1999). *AMOS 4.0*. Chicago, IL: SmallWaters Corporation.

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

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

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston: Houghton Mifflin.

Corno, L., Cronbach, L. J., Kupermintz, H., Lohman, D. F., Mandinach, E. B., Porteus, A. W., & Talbert, J. E. (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. Mahwah, NJ: Erlbaum

Detterman, D. K., & Daniel, M. H. (1989). Correlations of mental tests with each other and with cognitive variables are highest for low IQ groups. *Intelligence, 13*, 349-359.

Gustafsson, J. -E. (2002). Measurement from a hierarchical point of view. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 73-96). Mahwah, NJ: Erlbaum.

Horn, J. L. (1989). Models of intelligence. In R. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 29-73). Urbana: University of Illinois Press.

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

Jöreskog, K. G., & Sörbom, D. (1986). *LISREL VI: Analysis of linear structural relationships by maximum likelihood and least squares methods.* Mooresville, IN: Scientific Software.

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

Kaufman, A. S. (2000). Tests of intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (p. 445-476). New York: Cambridge University Press.

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.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, NJ: Prentice-Hall.

Lohman, D. F. (2000). Complex information processing and intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 285-340). Cambridge, UK: Cambridge University Press.

Lohman, D. F., & Hagen, E. (2001a). *Cognitive Abilities Test (Form 6).* Itasca, IL: Riverside.

Lohman, D. F., & Hagen, E. (2001b). *Cognitive Abilities Test (Form 6): Interpretive guide for school administrators.* Itasca, IL: Riverside.

Lohman, D. F., & Hagen, E. (2001c). *Cognitive Abilities Test (Form 6): Interpretive guide for teachers and counselors.* Itasca, IL: Riverside.

Lohman, D. F., & Hagen, E. (2002). *Cognitive Abilities Test (Form 6): Research handbook.* Itasca, IL: Riverside.

Lorge, I., & Thorndike, R. M. (1954). *Lorge-Thorndike Intelligence Tests.* Boston: Houghton Mifflin.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*, 519-530.

Messick, S. (1989). Validity. In R. Linn (Ed), *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53-61.

Steiger, J. H., & Lind, J. M. (1980). *Statistically-based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Sternberg, R. J. (1986). Toward a unified theory of human reasoning. *Intelligence, 10*, 281-314.

Tabachnick, B. D., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: HarperCollins.

Thorndike, R. M., & Hagen, E. (1971). *Cognitive Abilities Test: Multilevel Edition.* Boston: Houghton Mifflin.

Thorndike, R. M., & Hagen, E. (1974). *Cognitive Abilities Test (Multilevel Edition): Technical manual*. Boston: Houghton Mifflin.

Thorndike, R. M., & Hagen, E. (1987). *Cognitive Abilities Test (Form 4): Technical manual*. Chicago: Riverside.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs, 1*.

Wechsler, D. (1949*). Manual for the Wechsler Intelligence Scale for Children (WISC*). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children-Revised (WISC-R*). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children—Third Edition (WISC-III*). San Antonio, TX: The Psychological Corporation.