

Confirmatory Factor Analysis of the Iowa Tests of Basic Skills

Joseph J. Stevens

University of New Mexico

Confirmatory factor analysis was used to explore the internal structural validity of the Iowa Tests of Basic Skills (ITBS) using three samples: (a) the third-grade standardization sample, (b) a fourth-grade sample reported by Klein (1981), and (c) a new sample of third graders. Results showed that a four-factor structure provided a better representation of the relations among the 11 subtests than Klein's one-factor model, a two-factor model, or the three-factor model described in the ITBS manual. The four-factor model was refined using the new sample and was cross-validated using the other two samples. A hierarchical model with a single, second-order general achievement factor was also found to fit the data well. Decomposition of subtest variance into common, specific, and error components indicated that little specific variance is associated with several ITBS subtests, raising the question of whether interpretation of individual subtests is warranted.

The Iowa Tests of Basic Skills (ITBS) is one of the most widely used and highly regarded instruments for assessing student achievement. The initial instrument was developed in 1935 and has been succeeded by several updated forms (Lane, 1992). The instrument is available in three batteries of tests: the Early Primary Battery for grades Kindergarten through 1.9 (Levels 5 and 6), the Primary Battery for grades 1.7 through 3.5 (Levels 7 and 8), and the Multilevel Battery for grades 3 through 9 (Levels 9-14). Levels within each battery were constructed to generally correspond to the average achievement of children at the corresponding age (i.e., Level 9 to 9 years old).

The ITBS was designed to provide information that can be used in improving instruction (Riverside Publishing, 1982) and to "provide for the comprehensive measurement of growth in the fundamental skills: listening,

word analysis, vocabulary, reading, language, work-study, and mathematics" (Hieronymus & Hoover, 1986, p. 1). In addition to these general purposes, the manual lists nine more specific purposes of the instrument: for example, "To determine the developmental level of the pupil," "To diagnose specific qualitative strengths and weaknesses," and "To report performance in the basic skills to parents, pupils, and the general public in objective, meaningful terms" (Hieronymus & Hoover, 1986, p. 1). Score reporting is most commonly characterized by the use of individual subtest scores as well as composite scores consisting of averages across subtests.

Despite the care with which the ITBS has been developed, intensive investment in content development and specification, and a long tradition of research supporting particular applications of the instrument (Linn, 1989; Riverside Publishing, 1982), there is surprisingly little published research describing the factorial or internal construct validity of the ITBS. The manual for test administrators (Hieronymus & Hoover, 1986) provides partial information on unpublished factor-analytic studies using the standardization sample and suggests a three-factor structure. However, only two published studies appear in the literature (Klein, 1981; Martin & Dunbar, 1985), each reaching conflicting conclusions about ITBS structure.

Klein (1981) examined the structure of the ITBS to determine whether separate interpretation of the 11 subtests was warranted despite a high degree of subtest intercorrelations. Klein's sample consisted of 790 fourth-grade and 729 eighth-grade students from a large Southeastern suburban school district. In addition, analyses were performed on the standardization sample subtest correlations reported in the ITBS Preliminary Technical Summary (Houghton Mifflin, 1979). Klein used principal axis exploratory factor analysis (EFA) with oblique rotation to examine the underlying structure of the 11 ITBS subtests. For both the Southeastern and the standardization samples, the EFA results were characterized by a large initial eigenvalue with substantially smaller succeeding eigenvalues (e.g., fourth-grade, Southeastern sample: 7.18, .83, .59, .42, etc.; fourth-grade, standardization sample: 7.69, .63, .54, .36, etc.). On the basis of comparisons of the magnitude of the subtest communalities with the magnitude of the eigenvalues, Klein concluded that a single common factor best represented the standardization sample and two common factors should be inspected for the Southeastern sample.

Based on results of both the unrotated principal factors and the two-factor oblique solution, Klein (1981) found substantial redundancy and a high degree of interrelationship among the ITBS subtests. Klein concluded that there was "little evidence for the meaningful differentiation of the 11 subtest scores" and cautioned that

The intercorrelations among the subtests and subtotals seemed to be too high to produce meaningful profiles for use in individualized diagnosis...the usefulness of the ITBS in terms of both cost and time effectiveness should thus be scrutinized with great care. (p. 543)

Martin and Dunbar (1985) used hierarchical EFA to determine whether, in addition to a general achievement factor as suggested by Klein, additional factors were present. They analyzed a sample of 2,771 fifth-grade students randomly selected from the 1977 joint standardization sample of the ITBS and the Cognitive Abilities Test. Martin and Dunbar's analysis was based on parcels of subtest items rather than on the subtests themselves. Forty-eight parcels were formed from the individual subtest items using the test content specifications. Each of the 11 ITBS subtests was represented by no fewer than three and no more than six item parcels.

As in Klein's study, principal axis EFA with oblique rotation was used to analyze the relations among the 48 item parcels. The analysis extracted six principal factors with eigenvalues greater than 1.0, accounting for 62% of the variation in the original parcels. The first factor was interpreted as a Verbal Comprehension dimension and included parcels from the Reading, Vocabulary, and Language Usage subtests. The second factor was labeled Language Mechanics and included item parcels from the Capitalization, Punctuation, and Spelling subtests. Interpretation of the third factor was unclear, with the highest loadings associated with parcels of items from the Visual Materials, Math Concepts, and Math Problem-Solving subtests as well as some of the item parcels from the Reference Materials and Reading subtests. The fourth factor contained parcels of items from the Math Computation subtest. The fifth factor appeared to be related to the difficulty of item parcels and the sixth factor was not interpreted.

Intercorrelations of the six rotated factors ranged from .39 to .66 and were used as the basis for the second-order hierarchical analysis, which resulted in a single higher order factor. Results of the analysis showed that the general higher order factor dominated the solution, accounting for as much as 60% of the common variation among several of the parcel variables.

Martin and Dunbar (1985) concluded that their study provided "mixed support for the factorial validity of the ITBS" (p. 349). Although a strong general factor was found, they concluded that the importance of individual first-order factors had also been established. They interpreted their results as providing support for the convergent and discriminant validity of the Language and Mathematics subtests and concluded that the weakest support for ITBS structure was in terms of the Work-Study Composite comprised of the Reference Materials and Visual Materials subtests.

Although it is apparent from these two studies that there is a high degree of interrelationship among the subtests of the ITBS, it is unclear whether there is sufficient empirical evidence to justify the use and interpretation practices suggested in the ITBS manual and related publications. These practices recommend the interpretation of individual subtests, as well as composites of subtests. Although the results of Martin and Dunbar's study (1985) did demonstrate some degree of factorial complexity in the ITBS, their results must be interpreted cautiously given their use of constructed parcels of items rather than the subtest scores actually used for score report and interpretation.

In addition, there is some room for debating the conflicting interpretations resulting from the application of EFA in each of the described studies. In all reported results to date (i.e., Hieronymus & Hoover, 1986; Klein, 1981; Martin & Dunbar, 1985), only EFA methods have been used. The advantages of confirmatory factor analysis (CFA) over EFA for such purposes are well documented (Bagozzi, 1980; Bollen, 1989; Byrne, 1989; Gorsuch, 1983; Huba & Bentler, 1982; Marsh, 1987; Tanaka & Huba, 1984). They include the ability to formulate, define specifically, and test an a priori model; the ability to selectively specify or estimate particular model parameters; and the opportunity to directly test and compare the relative goodness of fit of competing models. These advantages appear to be particularly compelling in assessing the adequacy of published models of ITBS structure.

The purpose of this study, therefore, was to examine the alternative structural models proposed or suggested by Hieronymus and Hoover (1986), Martin and Dunbar (1985), and Klein (1981). The study also provided a more rigorous and direct evaluation of the relative goodness of fit of these models through application of confirmatory rather than exploratory factor analysis.

METHOD

Samples

Three samples were analyzed. Analysis of the first sample was based on the published correlation matrix for the third-grade (Level 9) standardization sample reported in the ITBS Manual (Hieronymus & Hoover, 1986). This sample was part of the 1984 national standardization of ITBS Form G. Standardization was based on a national probability sample which selected sampling units primarily on the basis of size of school district, region of the country, and socioeconomic characteristics. Ethnic composition of the standardization sample was reported as follows: Native American 1.2%, Asian 2.2%, Hispanic 8.6%, Black 22.0%, and White 65.5%. The sample was composed of 14,529 examinees.

The second sample was reported by Klein (1981) and was composed of 790 fourth-grade students who attended school in a large suburban school district in the Southeastern United States. ITBS Form 7 (Level 10) was administered during October of the 1979-1980 school year. Analyses reported here are based on the published correlation matrix of ITBS test scores reported in Klein (1981).

The third sample was composed of all third-grade students taking the ITBS Form J (Level 9) in a large suburban school district in the Southwestern United States. The ITBS was administered during late Spring, 1992. A total of 7,012 students took the ITBS. Three hundred-fifty students had at least one missing subtest score and were excluded from further study, resulting in a sample of 6,707. Of these, 3,376 (50.3%) were female, 3,326 (49.6%) were

male, and gender was not identified for 5 students. The ethnic composition of this sample was markedly different from the standardization sample. Two percent of the sample were identified as Asian, 3.6% as American Indian, 4.2% as Black, 38.7% as Hispanic, 48.0% as White, and 3.6% responded Other or did not respond.

ITBS

The ITBS Multilevel Battery is composed of 11 individual subtests: Vocabulary, Reading, Language Usage and Expression, Spelling, Capitalization, Punctuation, Visual Materials, Reference Materials, Mathematics Concepts, Mathematics Problem Solving, and Mathematics Computation. Internal consistency reliabilities (KR-20) for the individual subtests range from .80 to .91 for the third-grade standardization sample (Hieronymus & Hoover, 1986). In score reporting and discussions of score interpretation, the 11 subtests are grouped into five areas or subscales: Vocabulary, Reading, Language Skills, Work-Study, and Math Skills. Vocabulary and Reading are individual subtests, whereas the Language Total is composed of the average of Spelling, Capitalization, Punctuation, and Language Usage subtests. The Work-Study Total is made up of the average of the Visual Materials and Reference Materials subtests and the Mathematics Total represents the average of three subtests: Math Concepts, Math Problem-Solving, and Math Computation. In addition, two composite scores are provided: a Basic Composite composed of the average of Vocabulary, Reading, Spelling, and the Math Total and a Complete Composite composed of the average of the five subtest areas: Vocabulary, Reading, Language Total, Work-Study Total, and Math Total.

Structural Models

Four alternative conceptualizations of ITBS structure were tested using maximum likelihood CFA methods within LISREL 7 (Jöreskog & Sörbom, 1989). Correlation matrices reported by Klein (1981) and Hieronymus and Hoover (1986) were converted to covariance matrices for analysis. The first factor loading for each latent variable was set to 1.0 to establish a metric for each latent variable. Latent variables were allowed to correlate freely. The variances of subtest uniquenesses were estimated, but covariances of uniquenesses were fixed to zero. In addition to the four theoretical models of interest, a fifth null model (Model 0) was used to establish a baseline for model comparisons. More complex five- and six-factor models were considered but were not tested due to an inadequate number of subtests for each latent variable. Analyses of factor structure by examinee subgroup (e.g., gender, ethnicity) were not possible because only correlation matrices for total groups were available for the first two samples.

Model 1 is that suggested by Klein (1981), a single *General Achievement* factor which accounts for all 11 subtests (see Figure 1). Model 2 postulates two latent variables accounting for ITBS subtest performance: a *Language* factor and a *Mathematics* factor. Identification of the subtests with each factor was straightforward, based on subtest content, with the exception of the two study skills subtests, *Visual Materials* and *Reference Materials*. These subtests were associated with the *Mathematics* factor on the basis of results reported by Martin and Dunbar (1985).

Model 3 is the structure described in the ITBS Manual for Administrators (Hieronymus and Hoover, 1986) based on unpublished factor analytic studies using both varimax and oblique EFA. Model 3 was composed of three latent variables:

1. *Verbal/Reading*—Vocabulary, Reading, Language Usage and Expression, Spelling, Visual Materials, and Reference Materials subtests.
2. *Language Mechanics*—Capitalization and Punctuation subtests.
3. *Mathematics*—Mathematics Concepts, Mathematics Problem Solving, and Mathematics Computation subtests.

Analysis of Klein's fourth-grade sample included a minor variation, described in the manual for this grade level, in which the Spelling subtest loaded on the *Language Mechanics* factor rather than the *Verbal/Reading* factor.

Model 4 was suggested by structural features of the instrument, by the design and organization of score reporting and interpretive materials, and by inspection of the results of the six-factor analysis described by Martin and Dunbar (1985). Model 4 posits a four-factor representation composed of:

1. *Verbal Comprehension*—Vocabulary and Reading subtests.
2. *Language Mechanics*—Language Usage and Expression, Spelling, Capitalization, and Punctuation subtests.
3. *Study Skills*—Visual Materials and Reference Materials subtests.
4. *Mathematics*—Mathematics Concepts, Mathematics Problem Solving, and Mathematics Computation subtests (see Figure 1).

Replication and Cross-Validation

The availability of three independent samples for analysis allowed both replication and cross-validation of results. In evaluating the goodness of fit of the four competing structural models across samples, the intent was not cross-validation but independent testing of which of the four competing models was the best representation. These separate tests conducted in each sample were therefore viewed as replications.

Following choice of the best fitting of the competing models, refinement of the best model was undertaken. However, iterative, ex post facto model fitting of this type is subject to capitalization on chance and overfitting of the

model to the individual sample (see MacCallum, Roznowski, & Necowitz, 1992). To guard against this eventuality, the refined model was cross-validated. Model refinement was conducted only on the Southwestern sample (i.e., the calibration sample) and the other two samples were used as cross-validation samples.

To evaluate the degree of cross-validation, a series of nested invariance tests were conducted in which results for the calibration sample were applied to the cross-validation samples. The nested series of tests proceeded from less to more restrictive hypotheses of invariance across samples in four steps examining the equivalence of (a) model structure, (b) model structure and factor loadings, (c) model structure, factor loadings, and latent variable variances and covariances, and (d) model structure, factor loadings, latent variances/covariances, and variable uniquenesses. This hierarchy of tests covers the full range of strategies from full to partial cross-validation as described by Bandalos (1993), Cudeck and Browne (1983), or MacCallum, Roznowski, Mar, and Reith (1994).

Goodness of Fit

A large number of goodness-of-fit indices are available for the evaluation of structural equation models, but there is some disagreement as to their adequacy in assessing model fit (Bentler, 1990; Bollen, 1989; Browne & Cudeck, 1993; Marsh, Balla, & McDonald, 1988; Mulaik et al., 1989; Tanaka, 1993). A number of indices are influenced by sample size (Bentler, 1990; Bollen, 1989; Browne & Cudeck, 1989; LaDu & Tanaka, 1989; Marsh et al., 1988; Mulaik et al., 1989) and particular indices appear to be differentially suited as evaluations of particular conceptualizations of model fit (Tanaka, 1993) or for alternative purposes of model testing (Jöreskog, 1993).

The primary purpose of this study was to compare the relative goodness of fit of several competing models. Greatest concern was therefore in the relative ranking of the models rather than the absolute accuracy of an index as an indicator of model fit. In addition, many of the apparent differences in the performance of these indices appears to occur only in small to moderately sized samples (i.e., $N < 500$), whereas this study was characterized by extremely large samples. Although a number of indices were computed¹, for brevity only four indices are reported here: χ^2 , Jöreskog and Sörbom's Goodness-of-Fit Index (GFI), the Root Mean Square Residual (RMSR), and two forms of the Tucker-Lewis Index (TLI).

¹In addition to the indices reported here, the Comparative Fit Index (Bentler, 1990), the Akaike Information Criterion (Akaike, 1987), and the Corrected Akaike Information Criterion (Bozdogan, 1987) were also computed. In all instances, rankings of the competing models using these indices produced the same relative ordering of the models as that resulting from use of the TLI indices.

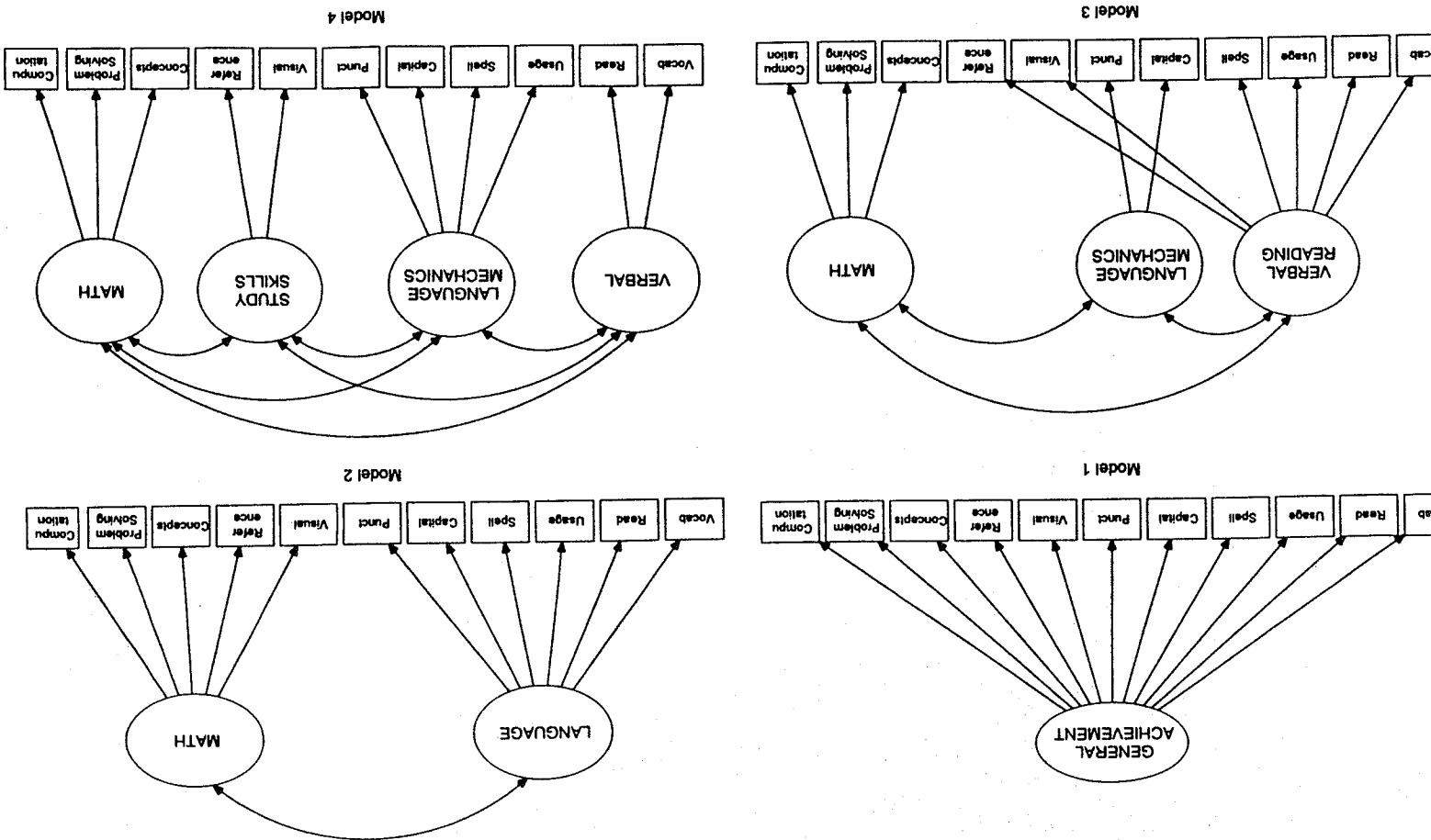


FIGURE 1 Four alternative models of ITBS structure.

The first form of the Tucker-Lewis Index (TLI_0) was:

$$TLI = (\chi^2_B / df_B - \chi^2_T / df_T) / (\chi^2_B / df_B - 1)$$

where χ^2_B / df_B is the ratio associated with a null, baseline model and χ^2_T / df_T is the ratio associated with a target model of interest (see Marsh et al., 1988; Tucker & Lewis, 1973). The second Tucker-Lewis Index (TLI_1) addresses concerns raised by Sobel and Bohnstedt (1985) regarding the appropriateness of the null model as a baseline for comparison in many model testing situations. They argued that the null model may be an unrealistic representation of the relations among variables and therefore is an inappropriate standard for comparison. In this study, theoretical interest was in the adequacy of a simple model (e.g., Klein's one-factor general achievement model) to fully account for subtest relations. To fully capture this interest, the TLI_1 index was computed using the one-factor model rather than the null as the baseline for comparison. Thus, the index describes the relative reduction in lack of fit of a more complex target model over the most parsimonious one-factor model.

RESULTS

Correlations with decimals omitted for the 11 ITBS subtests for the Southwestern sample appear in the Appendix. All subtest correlations were at least moderate in size, ranging from .499 to .799 and are generally similar in magnitude to those reported by Klein (1981) and by Hieronymus and Hoover (1986). For each sample, the null model (Model 0) and the four competing theoretical models were tested. Results of these analyses for the three samples appear in Table 1.

As would be expected given the large size of the samples, the observed χ^2 values were significant and quite large for all models tested. If one considers a model as only an approximation to reality (see Cudeck & Browne, 1983), then sufficiently large samples will always result in significance despite other indications of model fit. In this study, inspection of the other indices of goodness of fit supplied a more interpretable basis for comparing the alternative models.

The one-factor model provided moderate fits to the sample data, with GFI indices greater than .85 and RMSR values less than .05 for all samples. TLI_0 values ranged from .89 for Klein's sample to .93 for the standardization sample, indicating that the majority of the lack of fit observed in the null model was accounted for by the one-factor model.

Additional improvements in model fit occurred for the two-factor model. In all three samples, all goodness-of-fit indices improved in comparison to the one-factor model. GFI values were .91 or larger, values of RMSR were less than .04, and values of TLI_0 were .93 or larger. Computation of TLI_1 showed that, in the comparison of the two-factor model to the one-factor

TABLE 1
Summary of Alternative Models for ITBS Structure

Model	Sample		
	Southwestern ^a	Standardization ^b	Klein's ^c
Null (df = 55)			
χ^2	61,391	126,154	6,737
GFI	.191	.199	.206
RMSR	.594	.578	.567
One-factor (df = 44)			
χ^2	3,703	6,829	632
GFI	.895	.911	.850
RMSR	.036	.036	.048
TLI_0	.925	.933	.890
Two-factor (df = 43)			
χ^2	2,564	5,816	404
GFI	.925	.920	.909
RMSR	.030	.033	.038
TLI_0	.947	.941	.931
TLI_1	.295	.130	.372
Three-factor (df = 41)			
χ^2	1,918	4,101	400
GFI	.946	.946	.910
RMSR	.027	.029	.038
TLI_0	.959	.957	.928
TLI_1	.450	.357	.344
Four-factor (df = 38)			
χ^2	1,400	4,161	213
GFI	.960	.943	.954
RMSR	.023	.028	.027
TLI_0	.968	.953	.962
TLI_1	.569	.299	.656

Note. GFI = Goodness-of-fit index. RMSR = root mean square residual. TLI_1 = Tucker-Lewis index.

^aN = 6,707. ^bN = 14,529. ^cN = 790.

model, the smallest improvement in fit occurred for the standardization sample (.13) and the largest improvement for Klein's sample (.37).

Additional improvement in model fit was associated with the three-factor model for the Southwestern and standardization samples. For both samples, GFI values were .95, RMSR values were .03, and TLI_0 values were .96. TLI_1 values were .45 for the Southwestern sample and .36 for the standardization sample. Goodness of fit of Model 3 in Klein's sample was very similar to Model 2 with no substantial difference in the magnitude of indices for the two models.

Model 4 produced the best fit of the competing models for the Southwestern sample and Klein's sample. For these two samples, GFI values were larger than .95, RMSR values were smaller than .03, and TLI_0 values were

.97 and .96, respectively. TLI₁ values of .57 and .66 indicated that the majority of lack of fit for the one-factor model was accounted for by the four-factor model. Goodness-of-fit indicators for Model 4 in the standardization sample were similar, but slightly worse than those obtained for Model 3.

In summary, a unidimensional conceptualization of the ITBS (Model 1) produced the least acceptable fits for all three samples. The two-factor model, composed of correlated Language and Mathematics factors, fit better than a unidimensional representation for all samples. The three-factor model, composed of Verbal/Reading, Language, and Mathematics factors, provided additional improvements in goodness of fit and was the best fitting model for the standardization sample. The four-factor model provided the best fit for the Southwestern and Klein's samples and, in the standardization sample, produced comparable fit to the three-factor model.

Model Refinement

Following comparison of the alternative structural models, Model 4 results for the Southwestern sample were examined for possible refinements that would allow a better fit to the data. This was accomplished by relaxing particular parametric constraints individually as indicated by the largest modification index (see Byrne, 1989; Jöreskog, 1993). Examination of modification indices indicated that the largest improvement in model fit could be obtained by allowing the unique components (9%) of the Capitalization and Punctuation subtests to correlate. Following this alteration, the largest modification index involved freeing the Language Usage and Expression subtest to load on the Verbal factor as well as the Language factor. Although additional statistically significant modifications were present, they were not substantively interpretable and no further refinements were made.

Following these two modifications, goodness-of-fit indices for Model 4 in the Southwestern sample were: $\chi^2(36, N = 6,707) = 767.27$, adjusted goodness-of-fit index (AGFI) = .963, RMSR = .018, TLI₀ = .982, TLI₁ = .755. Standardized parameter estimates for the refined Model 4 appear in Table 2. All factor loadings were significant. It can be seen that all loadings of subtests on latent variables were positive and large in magnitude, with the exception of the split loading of the Language Usage and Expression subtest on the Verbal Comprehension and Language Mechanics factors. Intercorrelation of the four latent variables were also all significant and large in magnitude, ranging from .855 to .949 and suggesting a high degree of overlap among the four ITBS factors.

Hierarchical Structure

In addition to the models described previously, Martin and Dunbar's results (1985) suggest that another plausible conceptualization of ITBS structure is one involving a hierarchy in which first-order factors are mediated by a

TABLE 2
Standardized Parameter Estimates for the Refined Four-Factor Model

Subtest	Factor				SMC ^a
	Verbal Comprehension	Language Mechanics	Study Skills	Mathematics	
Factor Loadings					
Vocabulary	.878				.772
Reading	.908				.825
Language Usage	.373	.509			.737
Spelling		.815			.663
Capitalization		.806			.650
Punctuation		.782			.612
Visual Materials			.831		.690
Reference Materials			.851		.724
Math Concepts				.869	.756
Math Problem Solving				.866	.749
Math Computation				.740	.547
Factor Correlations					
Verbal Comprehension		.893	.923	.855	
Language Mechanics			.921	.882	
Study Skills				.949	

^aSMC = squared multiple correlation, a lower-bound estimate of subtest reliability.

second-order general achievement factor. This possibility is supported by the pattern of high factor correlations observed earlier. To test this conceptualization, Hierarchical Confirmatory Factor Analysis (HCFA) was used. The refined version of Model 4 was used as the representation of the first-order structure among the subtests. The HCFA model additionally specified a single second-order factor, General Achievement, associated with each of the four latent variables and accounting for the observed correlations among the latent variables.

Marsh (1987) described how the goodness of fit of a HCFA model can not exceed that of the corresponding first-order model. The goodness of fit of the hierarchical model can therefore be evaluated through comparison to two first-order models that provide limiting cases for the hierarchical model. The best fit and an upper limit for goodness of fit is provided by a first-order model in which all factor covariances are freely estimated (as in Model 4). An absolute lower limit of the goodness of fit of the hierarchical model is provided by a first-order model in which all factor covariances are fixed to zero.

The goodness of fit of Model 4 (with freely estimated factor covariances) was $\chi^2(36) = 767$, AGFI = .963, RMSR = .018, TLI₀ = .982, TLI₁ = .755, whereas the goodness of fit of the first order model that constrained all factor covariances to zero was $\chi^2(42) = 23,096$, AGFI = .373, RMSR = .515, TLI₀ = .508. The hypothesized hierarchical model provided a fit similar to Model

TABLE 4
Comparisons of the Calibration and Cross-Validation Samples

	χ^2	df	GFI	RMSR	TLI ₀	TLI ₁
Calibration: Southwestern sample	767	36	.980	.018	.982	.755
Cross-validation: Standardization sample						
Freely estimated	2,451	36	.969	.022	.971	.565
Λ fixed	3,118	45	.960	.027	.970	.557
Λ, Φ fixed	3,522	55	.954	.035	.973	.591
All parameters fixed	3,701	66	.954	.032	.976	.643
Cross-validation: Klein's sample						
Freely estimated	194	36	.959	.026	.964	.672
Λ fixed	215	45	.954	.031	.969	.717
Λ, Φ fixed	257	55	.943	.049	.970	.725
All parameters fixed	323	66	.932	.045	.968	.709

Note. GFI = Goodness-of-fit index. RMSR = root mean square residual. TLI = Tucker-Lewis index.

Cross-Validation

Refinement of Model 4 was conducted using only the Southwestern sample which served as a calibration sample in a nested series of cross-validation tests. The standardization and Klein's samples were used as cross-validation samples. In the least restrictive test, Model 4 was applied to the two cross-validation samples, but parameters were freely re-estimated in each sample (see Table 4). For both cross-validation samples, the structure of the refined model fit well and represented an improvement in fit over the original four-factor model for each sample (see Table 1). The next three tests introduced increasingly greater constraints over the fitted model by fixing parameters in the cross-validation samples to the values estimated for the calibration sample. Holding factor loadings (Λ Fixed), factor variances and covariances (Φ Fixed), and subtest uniquenesses (Θ ; Fixed) invariant introduced successively greater lack of fit as indicated by the obtained χ^2 values in each sample. However, even when all parameters were constrained, goodness-of-fit indices in the two cross-validation samples were remarkably good with GFI values greater than .93 and RMSR values less than .05. The reported TLI values for the cross-validation samples were also high but are somewhat inflated by increases in χ^2/df ratio due to the fixing of parameter estimates.

DISCUSSION

The primary purpose of this study was to evaluate competing models of ITBS structure using CFA and clarify the conflicting EFA results reported in previous studies. The results suggest that separable dimensions exist for the ITBS, contrary to some conclusions drawn by Klein (1981) and Martin and

4 with goodness-of-fit indices of $\chi^2(37) = 881$, AGFI = .959, RMSR = .019, TLI₀ = .978, TLI₁ = .725. Computation of Marsh and Hocevar's (1985) target coefficient, the ratio of the first-order chi-square to the second-order chi-square, provides an indication of the first order covariation accounted for by the second-order model. In the present case this ratio was 82%. Thus a substantial portion of the variation in the four-factor first-order model was accounted for by the general achievement factor. This was also reflected in the large magnitude of the standardized loadings of the first-order factors on the second-order, general achievement factor: .93 for Verbal Comprehension, .94 for Language Mechanics, .99 for Study Skills, and .94 for Mathematics.

Results of the HCFA analysis were also used to estimate the magnitude of particular variance components in the factor analytic model. Several authors have described the use of HCFA for this purpose—to estimate the proportion of variance associated with (a) common factors, (b) specific features of the measured variables, and (c) random measurement error of the variables (Jöreskog & Sörbom, 1989; Marsh & Grayson, 1994; Raffalovich & Bohrnstedt, 1987). This was accomplished by computing the diagonal elements of $(\Lambda\Gamma\Phi\Gamma\Lambda)$, $(\Lambda\Psi\Lambda)$, and (Θ) and relating the results to the total variance for each subtest using the completely standardized solution in LISREL 7 (see Jöreskog & Sörbom, 1989, pp. 162–163).

Proportions of each of the three variance components for each subtest are listed in Table 3. The proportion of common variance ranged from .48 for Math Computation to .72 for Reference Materials. The proportion of variance specific to each subtest was relatively small, with the largest specific components associated with the Vocabulary and Reading subtests and almost no specific variance associated with the two Study Skills subtests. Error variance ranged from .17 to .46. With the exception of the Reading subtest, the proportion of error variance associated with any particular subtest was larger than the reliable specific variance by a factor of at least two.

TABLE 3
Estimates of Common, Specific, and Error Variance Components of the ITBS Subtests

Subtest	Common	Specific	Error
Vocabulary	.66	.11	.23
Reading	.71	.12	.17
Spelling	.58	.08	.34
Capitalization	.58	.08	.34
Punctuation	.54	.08	.38
Language Usage	.68	.05	.24
Visual Materials	.68	.01	.31
Reference Materials	.72	.01	.27
Math Concepts	.67	.09	.24
Math Problem Solving	.66	.09	.25
Math Computation	.48	.06	.46

Dunbar (1985). Although there was substantial redundancy among the 11 subtests, a unidimensional model did not represent the relations among the subtests well. Use of the TLI₁ index showed that all competing models provided some improvement over the one-factor model. In fact, the single best-fitting model in two of the three samples was composed of four factors: Verbal Comprehension, Language Mechanics, Study Skills, and Mathematics. Following model refinement, this four-factor model provided substantially better fits than the competing models in all three samples.

Application of the hierarchical confirmatory analyses demonstrated that, although a four-factor structure was supported, a single, general achievement factor accounted well for the relations among the four factors. Loadings of the general achievement factor with the first-order factors were large and the majority of first-order variance was accounted for by the second-order factor in all samples.

The four-factor model refined in the Southwestern sample was cross-validated using the remaining two samples. MacCallum et al. (1994) argued, based on considerations of the sources of error in factor analysis (MacCallum & Tucker, 1991), that it may not be completely appropriate to cross-validate all model parameters as suggested by Cudeck and Browne (1983). Although factor loadings should not be influenced directly by sampling variation, the variances and covariances of latent and unique variables should be expected to vary from sample to sample. This suggests that a partial cross-validation strategy as described by MacCallum et al. (1994) may be most appropriate in many situations. In this study, both partial and complete constraint of model parameters across samples produced very little change in fit and, given the extremely large sample sizes used, there did not appear to be more than a small degree of degradation of model fit. These results support the generalizability of the four-factor model for third-grade examinees.

These results can all be interpreted in terms of evidence for the internal validity of the ITBS. Generally, the confirmatory analyses support the structure of the instrument. Discriminant validity was provided by the superiority of the four-factor model over the simpler models. Concurrent validity was provided by the strong loadings of subtests on their appropriate factor and by the observed relations of all subtests to a common general achievement factor.

As described by Messick (1989), however, validity must also be apparent in the use and interpretation of score information. This suggests that score reports, summaries, profiles, and interpretive materials should be supported by validity evidence and patterned after dimensions of an instrument that are demonstrably reliable and valid. The ITBS reports total scores and profile information for all 11 individual subtests; for Language Mechanics, Study Skills, and Mathematics composites; and for the Basic Composite and the Complete Composite. Use of all composite scores appears to be well supported by the results of this study.

Individual use and interpretation of subtest scores was not clearly supported by this study. There was little apparent differentiation among the individual subtests. Furthermore, results of the variance decomposition analysis indicated that there was little reliable variance not associated with the common factors for many of the subtests. Thus, report and interpretation of individual subtests for profile analysis, instructional diagnosis, and educational evaluation may not be warranted. If the individual subtests are to be used to represent separable facets of achievement, greater differentiation of the subtests appears to be needed.

ACKNOWLEDGMENTS

I thank Mike DuPuis and Carroll L. Hall for their help in making the data available, Patricia Clauser for her substantial efforts in preparing the data, and Randall Schumacker for helpful suggestions on cross-validation.

REFERENCES

- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52, 317-332.
- Bagozzi, R. P. (1980). *Causal models in marketing*. New York: Wiley.
- Bandalos, D. L. (1993). Factors influencing cross-validation of confirmatory factor analysis models. *Multivariate Behavioral Research*, 28, 351-374.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238-246.
- Bollen, K. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bozdogan, H. (1987). Model selection and Akaike's information criteria (AIC): The general theory and analytical extensions. *Psychometrika*, 52, 345-370.
- Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research*, 24, 445-455.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.
- Byrne, B. M. (1989). *A primer of LISREL: Basic applications and programming for confirmatory factor analytic models*. New York: Springer-Verlag.
- Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, 18, 147-167.
- Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Hieronymus, A. N., & Hoover, H. D. (1986). *Manual for school administrators, levels 5-14, ITBS Forms G/H*. Chicago: Riverside.
- Houghton Mifflin. (1979). *Preliminary technical manual* (2nd ed.). Iowa City, IA: Author.
- Huba, G. J., & Bentler, P. M. (1982). On the usefulness of latent trait modeling in testing theories of naturally occurring events. *Journal of Personality and Social Psychology*, 43, 604-611.
- Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294-316). Newbury Park, CA: Sage.
- Jöreskog, K. G., & Sörbom, D. (1989). *Lisrel 7: A guide to the program and applications* (2nd ed.). Chicago: SPSS.

- Klein, A. E. (1981). Redundancy in the Iowa Tests of Basic Skills. *Educational and Psychological Measurement*, 41, 537-544.
- LaDu, T. J., & Tanaka, J. S. (1989). Influence of sample size, estimation method, and model specification on goodness-of-fit assessments in structural equation models. *Journal of Applied Psychology*, 74, 625-635.
- Lane, S. (1992). Review of the Iowa Tests of Basic Skills, Form J. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 421-423). Lincoln, NE: University of Nebraska Press.
- Linn, R. L. (1989). Review of the Iowa Tests of Basic Skills, forms G and H. In J. C. Conoley & J. J. Kramer (Eds.), *The tenth mental measurements yearbook* (pp. 393-395). Lincoln, NE: University of Nebraska Press.
- MacCallum, R. C., Roznowski, M., Mar, C., & Reith, J. V. (1994). Alternative strategies for cross-validation of covariance structure models. *Multivariate Behavioral Research*, 29, 1-32.
- MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 111, 490-504.
- MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, 109, 502-511.
- Marsh, H. W. (1987). The hierarchical structure of self-concept and the application of hierarchical confirmatory factor analysis. *Journal of Educational Measurement*, 24, 1-39.
- Marsh, H. W., Balla, J., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 103, 391-410.
- Marsh, H. W., & Grayson, D. (1994). Longitudinal confirmatory factor analysis: Common, time-specific, item-specific, and residual-error components of variance. *Structural Equation Modeling*, 1, 116-145.
- Marsh, H. W., & Hocevar, D. (1985). The application of confirmatory factor analysis to the study of self-concept: First and higher order factor structures and their invariance across age groups. *Psychological Bulletin*, 97, 562-582.
- Martin, D. J., & Dunbar, S. B. (1985). Hierarchical factoring in a standardized achievement battery. *Educational and Psychological Measurement*, 45, 343-351.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: MacMillan.
- Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, 105, 430-445.
- Raffalovich, L. E., & Bohrnstedt, G. W. (1987). Common, specific, and error variance components of factor models: Estimation with longitudinal data. *Sociological Methods and Research*, 15, 385-405.
- Riverside Publishing. (1982). *Research that built the Iowa Tests of Basic Skills*. Chicago: Author.
- Sobel, M. E., & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. In N. B. Tuma (Ed.), *Sociological Methodology* (pp. 152-178). San Francisco: Jossey-Bass.
- Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10-39). Newbury Park, CA: Sage.
- Tanaka, J. S., & Huba, G. J. (1984). Confirmatory hierarchical factor analyses of psychological distress measures. *Journal of Personality and Social Psychology*, 46, 621-635.
- Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1-10.

APPENDIX
 ITBS Correlations, Means, and Standard Deviations for the Southwestern Sample (N = 6,707)

Subtest	1	2	3	4	5	6	7	8	9	10	11	M	SD
Vocabulary	1,000	799	1,000	675	621	645	664	631	723	1,000	664	104.8	14.3
Reading	1,000	799	1,000	675	621	645	664	631	723	1,000	664	103.9	17.2
Spelling	675	621	645	664	631	723	1,000	668	679	658	1,000	102.2	18.4
Capitalization	621	645	664	631	723	1,000	668	679	658	1,000	1,000	104.6	17.0
Punctuation	593	611	631	723	1,000	668	679	658	1,000	1,000	1,000	103.7	17.2
Language Usage	729	747	683	679	658	1,000	668	679	658	1,000	1,000	103.7	18.8
Visual Materials	660	697	590	600	605	668	679	658	1,000	1,000	1,000	104.8	15.7
Reference Materials	686	721	649	639	628	711	707	707	1,000	1,000	1,000	100.9	13.5
Math Concepts	662	678	610	618	616	666	713	685	1,000	1,000	1,000	104.0	13.8
Math Problem Solving	666	668	604	595	596	670	693	688	641	1,000	1,000	103.4	14.5
Math Computation	499	538	536	578	577	546	574	594	641	648	1,000	104.4	10.4

Note. Decimals omitted from correlations.