Psychological Assessment
1992, Vol. 4, No. 1, 102-112

1040-3590/92/$3.00

Review of Validity Research on the Stanford-Binet Intelligence Scale: Fourth Edition

Jeff Laurent and Mark Swerdlik
Illinois State University

Mary Ryburn
Unit 5 Public Schools
Normal, Illinois

The Stanford-Binet Intelligence Scale: Fourth Edition (SB:FE; Thorndike, Hagen, & Sattler, 1986) represents a significant departure from earlier versions of the scale. In the 5 years since its introduction into the field of intellectual assessment, a number of validity studies have been conducted with the SB:FE. The results from these construct and criterion-related validity studies suggest that the SB:FE provides as valid a measure of general mental ability as existing tests. Support for the 4 factors hypothesized by the authors of the SB:FE is weaker. Research suggests that the SB:FE is a 2-factor test (Verbal, Nonverbal) for ages 2 through 6 years and a 3-factor test (Verbal, Nonverbal, Memory) for ages 7 years and older. Studies also suggest that the SB:FE can distinguish between groups of youngsters with differing intellectual abilities (e.g., mentally handicapped, gifted, neurologically impaired) and that the test correlates highly with scores on achievement tests. On the basis of validity information, recommendations for the use of the SB:FE are made.

In its various revisions since 1916 the Stanford-Binet Intelligence Scale has been a mainstay of assessors for decades. It was the first published intelligence test to provide specific administration and scoring procedures. It was also the first American test to use the concept of the intelligence quotient (IQ). As time has passed, revisions in the Wechsler scales and the advent of new measures of intelligence (e.g., Kaufman Assessment Battery for Children, Kaufman & Kaufman, 1983; Woodcock-Johnson Tests of Cognitive Ability, Woodcock & Johnson, 1977) have replaced the Stanford-Binet as the instrument of choice in assessing the intellectual functioning of children (Aiken, 1987; Freides, 1972; Lubin, Larsen, & Matarazzo, 1984). Therefore, the revision of the Stanford-Binet in 1986 was timely.

The Stanford-Binet Intelligence Scale: Fourth Edition (SB:FE; R. L. Thorndike, Hagen, & Sattler, 1986a) represents a significant departure from previous editions of the measure. Some continuity in the types of items included on the test exists between the SB:FE and earlier versions of the instrument. More items have been added to existing tasks, however, and six new types of items are included in the most recent revision. In addition, items have been grouped to form 15 subtests. Administration and scoring changes also are contained in this revision. Perhaps the most significant departures from earlier versions of the Stanford-Binet include a well-defined theoretical orientation, the change from an age scale to a point scale, updated norms, and suggestions for abbreviated batteries.

The dramatic changes in the most recent version of the Stanford-Binet led to a frenzy of activity. Practitioners and researchers have evaluated the SB:FE, comparing it with its immediate predecessor, Form L-M, and other existing measures of intelligence. The result of this scrutiny has been an body of literature that includes test reviews and critiques (e.g., Davis, 1989; Glutting, 1989; Reynolds, 1987), surveys of use (e.g., Chattin & Bracken, 1989; Hanson, 1989; Heath, 1988), and validity studies. The reviews and surveys provide information about users' reactions to the SB:FE, whereas the validity studies provide a more empirical basis for evaluating the SB:FE. This article reviews the validity studies that have been conducted with the SB:FE, in an attempt to better understand the nature of this revised measure.

Correspondence concerning this article should be addressed to Jeff Laurent, Department of Psychology, Illinois State University, Normal, Illinois 61761.

Validity Evidence

The validity of a test is judged by determining whether it actually measures what it proposes to measure. Most measurement texts discuss three types of validity: content, criterion-related, and construct. In this review, the focus is on the criterion-related and construct validity of the SB:FE. Criterion-related validity examines the relationship between scores on a test and some criterion that is measured either at the same time (i.e., concurrent criterion-related validity) or at some future time (i.e., predictive criterion-related validity). Construct validity refers to the extent to which a test measures the theoretical construct or trait it was intended to measure.

Factor-Analytic Studies: A Measure of Construct Validity

R. L. Thorndike et al. (1986a) were guided by Horn and Cattell's (1966; Cattell, 1963) hierarchical model of cognitive abilities when developing the SB:FE (see Figure 1). The hierarchical model on which the SB:FE was based consists of a general reasoning factor, g, at the top level; three broad factors at the sec-

ond level (i.e., Crystallized Abilities, Fluid-Analytic Abilities, Short-Term Memory); and a third level of more specific factors (i.e., Verbal Reasoning, Quantitative Reasoning, Abstract/Visual Reasoning). The existence of the more specific factors of the third level could be viewed as support for the Crystallized Abilities and Fluid-Analytic Abilities factors of the second level. Therefore, if factors corresponding to Verbal Reasoning, Quantitative Reasoning, Abstract/Visual Reasoning, and Short-Term Memory were found as a result of factor analysis, the theoretical model underlying the SB:FE would be supported.

Exploratory and confirmatory factor-analytic studies have been conducted on the SB:FE. Simply stated, in exploratory factor analysis the researcher allows the data to define the factor structure. This technique is generally used when the underlying structure of a data set is unknown. The rationale behind such an approach is that the analysis will result in a factor structure that can be compared to that on which a test may have been developed. Ideally, the actual factor structure is consistent with that theorized by the test developers. When this occurs, those using the test can feel confident that the test is measuring what it was intended to measure. If a factor structure emerges that is not consistent with that suggested by the theory on which a test was developed, however, then the validity of the test is questioned. Because the researcher is responsible for designating certain parameters for the analysis (e.g., eigenvalues of greater than 1 to enter the analysis, varimax rotation, number of factors) and subjectively interprets factors, the process is not as objective as it might be.

Confirmatory factor analysis takes a more objective approach to factor analysis. In this approach, the researcher makes explicit statements along a number of dimensions (e.g., the number of common factors, the variances and covariances among the common factors, the relationships among observed variables, and latent factors; Long, 1983). This specified factor model, which can be theory-bound, is then compared with the factor structure that emerges when actual data are analyzed, to determine the "goodness of fit."

The studies cited in the *Technical Manual* (R. L. Thorndike et al., 1986b) provide a starting point for examining the construct validity of the SB:FE. The *Technical Manual* reports the results of an unspecified "variant" of a confirmatory factor analysis that was conducted using the standardization sample

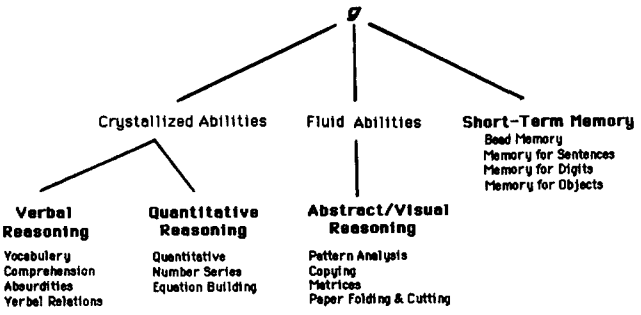

Figure 1. Hierarchical model of the cognitive abilities measured by the Stanford-Binet: Fourth Edition.

(R. L. Thorndike et al., 1986b). The purpose of such a factor analysis was to determine whether the theory on which the revision was built was supported when the test was administered to the standardization sample.

The factor analysis across all ages of the standardization sample reported in the *Technical Manual* did support a general factor, g. In addition, the Short-Term Memory factor of the second level, and the Verbal Reasoning, Quantitative Reasoning, and Abstract/Visual Reasoning factors of the third level emerged. Further factor analyses performed across three age ranges of the standardization sample (i.e., 2 through 6 years, 7 through 11 years, 12 through 18–23 years) and reported in the *Technical Manual* also found a general factor. Support for the other factors was modest, however, and varied across age groups. For example, support was found in the 2 through 6 years age group for Verbal and Abstract/Visual Reasoning factors but not for a Quantitative or Short-Term Memory factor. Verbal Reasoning, Short-Term Memory, and Abstract/Visual Reasoning factors were discovered for the age group 7 through 11 years; a Quantitative Reasoning factor was not found with this group. All four factors emerged for the 12 through 18–23 years age group. The *Technical Manual* provides evidence that substantiates the theoretical construct on which the SB:FE was built, the existence of a general factor, g, and Verbal Reasoning, Quantitative Reasoning, Abstract/Visual Reasoning, and Short-Term Memory factors. However, the evidence varies across age levels. The fact that the *Technical Manual* does not contain the goodness-of-fit index, the adjusted goodness-of-fit index, and the root mean square residual for the different confirmatory factor analyses makes it difficult to evaluate the results.

Subsequent exploratory factor analyses using the standardization sample (Reynolds, Kamphaus, & Rosenthal, 1988; Sattler, 1988) have reported results that differ from those presented in the *Technical Manual.* Generally, these principal-component analyses have supported the existence of a general factor, g, across all age ranges. There was, however, no support for a four-factor solution at any of the age ranges. The principal-component analysis conducted by Reynolds et al. (1988) found that two factors were produced at 10 age groups, and three factors were produced only at age 17 years. At best, the two-factor solution for children ages 2 and 3 years parallels the broad Crystallized Abilities and Fluid Ability factors hypothesized by the theoretical model. For most other age levels, Reynolds et al. suggested that Verbal Reasoning and Analytic/Sequential Processing best describe the factors. Reynolds et al. (1988) concluded, "At no age does the structure seem at all consonant with the structure hypothesized by Thorndike, Hagen, and Sattler" (p. 56).

Sattler's (1988) principal-components analysis supported the existence of two factors for the 2 through 6 years age group: Verbal Comprehension and Nonverbal Reasoning/Visualization, which roughly corresponded to the Verbal and Abstract/Visual Reasoning factors reported in the *Technical Manual.* However, Sattler's analysis found that a three-factor approach best characterized both the 7 through 11 years and 12 through 18–23 years age groups. These three factors, Verbal Comprehension, Nonverbal Reasoning/Visualization, and Memory, represent a departure from the results reported in the *Technical*

Manual, especially for the 12 through 18–23 years age group. The results of Sattler's principal-components analysis call into question the validity of the theoretical model on which the SB:FE was developed, especially as it relates to a Quantitative Reasoning component.

The factor-analytic work of Reynolds et al. (1988) and Sattler (1988) represented initial responses to the publication of the SB:FE. Two subsequent exploratory factor-analytic studies (i.e., Boyle, 1989; R. M. Thorndike, 1990) have profited methodologically from the opportunity to examine these earlier studies. Boyle has presented a persuasive argument for the utility of using an oblique simple structure rotational technique rather than using the traditional a priori decision to use an orthogonal rotation. Using an iterative principal factoring technique and rotating the extracted factors to an oblique simple structure by using the SPSS direct Oblimin procedure, Boyle found that a four-factor solution (Abstract/Visual Reasoning, Verbal Reasoning, Short-Term Memory, and Quantitative Reasoning) consistent with that proposed by R. L. Thorndike et al. (1986a) best described the SB:FE.

R. M. Thorndike (1990) provided a review of the exploratory factor-analytic studies on the SB:FE, highlighting their weaknesses. The most noticeable criticisms were leveled at the studies conducted by Reynolds et al. (1988) and Boyle (1989). R. M. Thorndike criticized the Reynold et al. factor analysis because he felt it provided nothing more than a "computer center default solution" (p. 418) to the standardization data. R. M. Thorndike was harsh in his criticism, because he later stated the importance of "letting the data speak" (p. 420). Reynolds et al. clearly were letting the data speak but perhaps in a more rigid fashion than that with which R. M. Thorndike was comfortable. It would seem more productive to examine what the data indicate within a number of different parameters than to criticize a study over a philosophical point. R. M. Thorndike's criticism of the Boyle (1989) study is an important one. By including the four area scores and the Test Composite score along with the 15 individual subtests in the factor analysis, Boyle increased the likelihood of finding the four factors hypothesized by R. L. Thorndike et al. (1986a, 1986b) because of the phenomenon of linear dependence. In other words, the subtests contribute to the area scores, and these in turn result in the Test Composite, so the data are not independent of one another, strengthening the four factors and the probability that they will emerge.

In addition to reviewing the existing exploratory factor analyses that have been done with the SB:FE, R. M. Thorndike (1990) conducted his own factor analysis using the standardization data in an attempt to address the shortcomings of previous analyses. Thorndike used a principal-axis factoring of correlation matrices with squared multiple correlations in the diagonal as initial communality estimates, and then used the iterations for the communalities in an oblique rotation. The results of this analysis and a second analysis that used the matrices of median correlations for the age groups used by R. L. Thorndike et al. (1986b) were similar. These analyses yielded a two-factor solution (i.e., Verbal, Nonverbal) for the 2- through 6-year-old group, and a three-factor solution (i.e., Verbal Ability, Abstract/Visual, Memory) for individuals from age 7 through 23 years. The Abstract/Visual factor included subtests from the Quantitative Reasoning and Abstract/Visual Reasoning areas. This combination is inconsistent with the second level of the hierarchy proposed by R. L. Thorndike et al. (1986a, 1986b), in which the Quantitative Reasoning subtests are associated with Crystallized Abilities and the Abstract/Visual Reasoning subtests are associated with Fluid-Analytic Abilities. The results from R. M. Thorndike's factor analyses are very similar to those of Sattler (1988) and of R. L. Thorndike et al. (1986b) with the 7- through 11-year-old group.

Several confirmatory factor-analytic studies of the SB:FE have been undertaken since its publication (Keith, Cool, Novak, White, & Pottebaum, 1988; Kline, 1989; Ownby & Carmin, 1988). These confirmatory factor analyses have used the intercorrelation matrices provided by R. L. Thorndike et al. (1986b) in the *Technical Manual* for the standardization sample and some version of the LISREL VI computer program (Jöreskog & Sörbom, 1986) to test the four-factor theory underlying the SB:FE (i.e., Verbal Reasoning, Quantitative Reasoning, Abstract/Visual Reasoning, Short-Term Memory).

Three statistics are commonly used to determine the goodness of fit in confirmatory factor analysis: the goodness-of-fit index, the adjusted goodness-of-fit index, and the root mean square residual (Cole, 1987). A goodness-of-fit index of greater than .90, an adjusted goodness-of-fit index of greater than .80, and a root mean square residual of less than .10 indicate that the data fit a specified model well (Cole, 1987). When using the entire standardization sample and a very strict factor structure that allowed each subtest to load on only one factor, Keith et al. (1988) reported an adjusted goodness-of-fit index of .879 and a root mean square residual correlation of .044. When a "relaxed" factor structure was examined, the goodness-of-fit statistics improved to .904 and .037, respectively. This relaxed factor structure allowed for additional factor loadings consistent with the constructs of the SB:FE. For instance, Memory for Sentences was allowed to load on Verbal Reasoning in addition to Short-Term Memory. Confirmatory factor analyses also were conducted on subsets of data that were designed to represent three populations, preschool (ages 2–6 years), elementary (ages 7–11 years), and adolescent and adult (ages 12–23 years). Generally, the results for these three subgroups paralleled those for the entire standardization sample. The adjusted goodness-of-fit indexes all exceeded .80, and the root mean square residuals were less than .10. The results of the confirmatory factor analyses support the underlying theory of the test (i.e., the existence of four factors corresponding to Verbal Reasoning, Quantitative Reasoning, Abstract/Visual Reasoning, and Short-Term Memory). The high correlations between the four factors suggest that the test is also a measure of g. An additional analysis, however, conducted by Keith et al. in which a g-only model provided a significantly worse fit to the data suggests that the SB:FE should not be viewed as strictly a measure of g.

Although the Keith et al. (1988) results appear to support the theoretical underpinnings on which the SB:FE was based, there were some important inconsistencies in the model. For example, the theoretical model states that the Verbal and Quantitative Reasoning factors represent crystallized intelligence and that the Abstract/Visual Reasoning factor represents fluid intelligence. One would expect that the Verbal and Quantitative Reasoning factors would correlate more highly with one another than with any other factor. The results of the confirma-

tory factor analyses, however, consistently reported correlations between the Quantitative and Abstract/Visual Reasoning factors that exceeded those between the Verbal and Quantitative Reasoning factors. Therefore, the results of the Keith et al. study do not support the second level of the hierarchical model of intelligence presented by R. L. Thorndike et al. (1986a, 1986b). Additionally, in the preschool group the Short-Term Memory factor correlated perfectly ($r = 1.0$) with both the Verbal and Abstract/Visual Reasoning factors, suggesting that a Short-Term Memory factor did not exist for this group. A subsequent analysis with this group found that the best goodness-of-fit statistics were obtained with a "no-memory" model. The improved goodness-of-fit statistics, which resulted when a relaxed model was used, suggest that some subtests (e.g., Bead Memory, Memory for Sentences, Absurdities) should be considered with factors other than those on which they currently are associated. The Keith et al. study can be viewed as providing mixed support for the construct validity of the SB:FE.

A second confirmatory factor analysis using four models was conducted by Ownby and Carmin (1988). Model I was a g-only model similar to that examined by Keith et al. (1988). Model II was a two-factor model in which Verbal, Abstract/Visual, and Short-Term Memory subtests with verbal content formed a Verbal factor and nonverbal subtests formed a Nonverbal factor. This model was based on the dichotomy that exists among the Wechsler scales. A three-factor structure, Verbal, Nonverbal, and Memory, composed Model III. Finally, Model IV reflected the model postulated by R. L. Thorndike et al. (1986a), which consisted of the Verbal, Quantitative, and Abstract/Visual Reasoning factors along with the Short-Term Memory factor. In addition to the entire standardization sample, Ownby and Carmin conducted confirmatory factor analyses on the four models for 2-, 4-, 6-, 8-, and 10-year-olds. Although the adjusted goodness-of-fit indexes for all models with each sample exceeded .80, Model IV provided the best fit for the majority of the data. The exception to this finding occurred with the 2-year-old sample, in which Model II provided the best fit. Model II also provided a strong challenge to Model IV among 6-year-olds. There was support for g, but as was the case in the Keith et al. (1988) study, the g-only model did not provide the best fit for the data. With the exception of the 2-year-old sample, the results support the theorized four-factor structure on which the SB:FE was based. Unfortunately, Ownby and Carmin do not provide the root mean square residuals or correlations among the first-order latent factors (i.e., Verbal Reasoning, Quantitative Reasoning, Abstract/Visual Reasoning, Short-Term Memory) that would allow further analyses of their data. Although unspecified, it is assumed that strict models were used that allowed subtests to load on one factor only. It would have been interesting to have seen how subtests would have loaded had relaxed models also been used.

Finally, Kline (1989) conducted confirmatory factor analyses to determine whether R. L. Thorndike et al.'s (1986a, 1986b) four-factor model or Sattler's (1988) two- or three-factor models provided the best goodness of fit for the standardization sample. Similar to Ownby and Carmin (1988), Kline did not specify whether a strict or more relaxed model was used in his analyses. He also did not conduct an analysis of the standardization sample as a whole. Instead, he performed analyses on each of the individual age groups included on the SB:FE. It is noteworthy that the adjusted goodness-of-fit indexes obtained by Kline (1989) for the four-factor model were lower than those obtained by Ownby and Carmin's (1988) Model IV for the same age samples, 2-, 4-, 6-, 8-, and 10-year-olds. Because both studies lack discussion of the parameters entered into the LISREL VI computer program, it is difficult to comment on these differences. Kline also reported several goodness-of-fit and adjusted goodness-of-fit indexes that exceeded the cutoffs suggested by Cole (1987). All of the root mean square residuals exceeded Cole's "less than .10" criterion. These results contradict those of the earlier reported confirmatory factor-analytic studies (Keith et al., 1988; Ownby & Carmin, 1988) and suggest that the four-factor model may not be supported by the standardization data. On the other hand, the confirmatory factor analyses using Sattler's (1988) two-factor model for each age group from 2 through 6 years and three-factor model for each age group from 7 through 18–23 years all resulted in goodness-of-fit indexes, adjusted goodness-of-fit indexes, and root mean square residuals that met Cole's criteria. Kline's results suggest that Sattler's (1988) two- and three-factor models are a better fit for the standardization data and more accurately reflect the structure of the SB:FE than does the four-factor model expounded by R. L. Thorndike et al. (1986a, 1986b).

In general, factor-analytic evidence in support of the SB:FE has been mixed, depending on the type of factor analysis conducted. With the exception of Boyle's (1989) study, which suffers from methodological problems, exploratory factor analyses have not supported the four-factor model on which the SB:FE was developed and suggest alternative factor structures. Unfortunately, several of the confirmatory factor-analytic studies lack the necessary information (e.g., root mean square residuals) for the reader to thoroughly evaluate their findings. Nevertheless, with the exception of Kline (1989), confirmatory factor analyses have generally supported the four-factor model. Significant caveats exist, however, concerning the validity of this model with children below age 6 years. Among this group of children, the Short-Term Memory factor does not appear to be valid (Keith et al., 1988; Ownby & Carmin, 1988). The findings from the Keith et al. study also call into question the second level of the R. L. Thorndike et al. (1986a, 1986b) hierarchical model and the notion that crystallized intelligence is measured by the Verbal Reasoning and Quantitative Reasoning factors. The Quantitative Reasoning factor was more highly correlated with the Abstract/Visual Reasoning factor associated with fluid-analytic intelligence.

Beyond the question of the validity of the four-factor model, each of the factor-analytic studies has supported the notion that the SB:FE is a measure of g. Confirmatory factor-analytic studies suggest more strongly than do exploratory factor analytic studies that g alone does not adequately describe what is being measured by the SB:FE.

The SB:FE and Other Measures of Intelligence: An Examination of Criterion-Related Validity

When a new or revised test is released, a flurry of studies appear in professional journals comparing it with existing instruments that purport to measure the same thing. Predictably,

this has been the response to the publication of the SB:FE. Several studies have compared the SB:FE with other measures of intelligence. It is important to remember when evaluating these studies that test authors may define *intelligence* differently. As a result, the criterion for comparison may not be consistent. Nevertheless, these comparisons provide insight into the validity of the SB:FE as a traditional measure of intelligence.

Again, the *Technical Manual* (R. L. Thorndike et al., 1986b) provides a starting point for examining the criterion-related validity of the SB:FE. Table 1 summarizes comparisons between the SB:FE Test Composite score and the equivalent full scale score of other intelligence measures reported in the *Technical Manual.* The global scores contained in Table 1 are the easiest to use in comparisons, because they all measure general intelligence. In addition, the full scale scores of the various instruments tend to be the most reliable and valid measures of intelligence. Therefore, if one is seeking the best measure of criterion-related validity, comparisons between full scale scores will provide this.

When examining Table 1, several methodological and psychometric considerations must be kept in mind. For instance, the socioeconomic status of the nonexceptional samples was higher than that of the general population (R. L. Thorndike et al., 1986b). In addition, the various studies using exceptional individuals relied on samples that were identified by school districts. School districts vary in their definitions and their reliance on test scores in making placement decisions. This selection process may have affected the results obtained in some of the studies. Also, most studies did not use a counterbalanced design. In most cases one test was administered to the entire sample before the other test. In some studies a considerable amount of time passed (i.e., 12–15 weeks) before the administration of the second measure. Several studies of exceptional individuals must be viewed cautiously because of small sample sizes. Restriction of range was a problem with several of the studies using exceptional samples, especially those with gifted and mentally retarded individuals. Studies comparing the SB:FE and WAIS–R represent a unique situation in which the former presents a better floor and the latter a better ceiling, important considerations in studies using mentally retarded and gifted populations, respectively.

Table 1
Correlations Between the Stanford-Binet Intelligence Scale: Fourth Edition Test Composite Score and Full Scale Scores of Other Intelligence Tests Reported in the Technical Manual

Criterion test	Sample	*n*	Mean age (in months)	*r*
SB:L-M	Nonexceptional	139	83	.81
SB:L-M	Gifted	82	88	.27
SB:L-M	Learning disabled	14	100	.79
SB:L-M	Mentally retarded	22	143	.91
WISC–R	Nonexceptional	205	113	.83
WISC–R	Gifted	19	155	.69
WISC–R	Learning disabled	90	132	.87
WISC–R	Mentally retarded	61	167	.66
K–ABC	Nonexceptional	175	84	.89
K–ABC	Learning disabled	30	107	.66
WAIS–R	Nonexceptional	47	233	.91
WAIS–R	Mentally retarded	21	234	.79
WPPSI	Nonexceptional	75	66	.80

Note. SB:L-M = Stanford-Binet Intelligence Scale: Form L-M; WISC–R = Wechsler Intelligence Scale for Children—Revised; K–ABC = Kaufman Assessment Battery for Children; WAIS–R = Wechsler Adult Intelligence Scale—Revised; WPPSI = Wechsler Preschool and Primary Scale of Intelligence.

Keeping in mind these potential limitations, one can evaluate the criterion-related validity of the SB:FE on the basis of the studies cited in the *Technical Manual.* One can use an arbitrary, conservative yardstick and view coefficients of .90 and greater as *very good,* coefficients of .80 and greater as *good,* coefficients of .70 and greater as *fair,* and coefficients of less than .70 as suggesting *poor* criterion-related validity. Using this standard, most studies cited in the *Technical Manual* suggest good to very good criterion-related validity for the SB:FE when compared with other widely used measures of intelligence. This was especially true of studies using nonexceptional samples. The ability of the samples in these studies to approximate that of the general population (i.e., the normal distribution) may explain these positive results.

Generally, there were wide variations when examining the validity coefficients for groups of exceptional individuals. The coefficients ranged from poor to good (i.e., .66–.87) for groups of learning disabled individuals and poor to very good (i.e., .66–.91) for groups of mentally retarded individuals. The validity coefficients resulting from the studies of gifted children were consistently poor. This was particularly true in the study of the SB:L-M and the SB:FE. The results from this study suggest that the revision is different from its predecessor: Form L-M relied heavily on verbal abilities at higher levels, whereas the SB:FE incorporates several elements. These differences may be especially noticeable when assessing children identified as gifted because verbal skills traditionally have been emphasized in this assessment. Poor validity coefficients for the gifted and other samples also may reflect a restricted range of scores in some of these studies. Potential ceiling effects, although not reported, may have occurred in the studies of gifted children using the SB:FE and the WISC–R, depending on their age. The older the gifted child being administered the WISC–R, the more likely it may be that there will not be enough difficult items to adequately measure intellectual ability. This is not as great a problem when using the SB:FE because of the broad age range covered by the test. Similarly, a lack of an adequate number of floor items may have occurred in the study of the performance of mentally retarded individuals on the SB:FE and WAIS–R.

Table 2 summarizes comparisons between the SB:FE Test Composite score and the equivalent full scale score of other intelligence measures that have been conducted since the publication of the SB:FE. Generally, these studies have addressed several of the methodological shortcomings of the studies reported in the *Technical Manual.* Although there were a few exceptions (Carvajal & Weyand, 1986; Goh & Cordoni, 1989; Lukens, 1988; Richardson, 1988), most used a counterbalanced design and administered tests within a few days or weeks of one another.

Approximately one half of the validity coefficients reported

Table 2
Correlations Between the Stanford-Binet Intelligence Scale: Fourth Edition Test Composite Score and Full Scale Scores of Other Intelligence Tests

Criterion test	Sample	*n*	Mean age (in months)	*r*
SB:L-M				
Clark et al. (1987)	Normal	47	63	.53
Hartwig et al. (1987)[a]	Normal	30	136	.72
Krohn & Lamp (1989)	Head start	89	59	.69
Livesay & Mealor (1987a)	Gifted[b]	120	82	.64
Lukens (1988)[a]	Trainable mentally retarded	31	201	.86
McCall et al. (1989)	Gifted	32	101	.21
WISC-R				
Carvajal & Weyand (1986)[a]	Normal	23	—	.78
Livesay & Mealor (1987b)	Gifted[b]	166	124	.55
Phelps (1989)	Gifted	48	139	.39
Phelps et al. (1988)	Learning disabled	35	143	.92
Rothlisberg (1987)	Normal	32	93	.77
Swerdlik & Ryburn (1989)	Referred	26	113	.86
K-ABC				
Hayden et al. (1988)[a]	Gifted	32	112	.70
Krohn & Lamp (1989)	Head start	89	59	.86
Rothlisberg et al. (1990)	Referred	40	80	.85
Smith & Bauer (1989)	Normal	30	59	.57[c]
Smith et al. (1989)	Learning disabled	18	125	.74[c]
WAIS-R				
Carvajal et al. (1987)	College	32	237	.91
Goh & Cordoni (1989)[a]	College learning disabled	38	223	.68
Richardson (1988)	Mentally retarded	35	—	.66
WPPSI				
Carvajal et al. (1988)	Normal	20	—	.59

Note. SB:L-M = Stanford-Binet Intelligence Scale: Form L-M; WISC-R = Wechsler Intelligence Scale for Children—Revised; K-ABC = Kaufman Assessment Battery for Children; WAIS-R = Wechsler Adult Intelligence Scale—Revised; WPPSI = Wechsler Preschool and Primary Scale of Intelligence.
[a] Abbreviated battery of the Stanford-Binet: Fourth Edition was used. [b] Children referred for a gifted program. [c] Correlation coefficient corrected for restricted range.

in Table 2 equaled or exceeded .70, suggesting fair to very good criterion-related validity between the SB:FE and other existing measures of intelligence. Most of the studies with validity coefficients that fell within the good (i.e., $r \geq .80$) to very good range (i.e., $r \geq .90$) used samples that had been referred for special education evaluation (e.g., Rothlisberg, McIntosh, & Dodge, 1990; Swerdlik & Ryburn, 1989), placed in Head Start programs (e.g., Krohn & Lamp, 1989), or were drawn from a general college population (e.g., Carvajal, Gerber, Hewes, & Weaver, 1987). Subjects included in these samples may have possessed a broader range of abilities or demonstrated more variability as a group, and therefore more closely approximated a normal distribution. Similarly, those studies that reported fair criterion-related validity results tended to consist of nonexceptional samples (Carvajal & Weyand, 1986; Hartwig, Sapp, & Clayton, 1987; Rothlisberg, 1987). In general, the highest validity coefficients were associated with studies whose samples demonstrated more variability as a group. Four studies that reported validity coefficients of greater than .70 did not fit this general pattern. Three (Hayden, Furlong, & Linnemeyer, 1988; Phelps, Bell, & Scott, 1988; Smith, St. Martin, & Lyon, 1989) used samples whose ages fell within the middle of the range covered by the tests, thus providing adequate floor and ceiling items. The other study (Lukens, 1988) used an older sample (mean age = 16 years) of individuals with moderate mental retardation. The high validity coefficient suggests that both the SB:FE and SB:L-M provide a sufficient number of floor items to assess the cognitive abilities of low-functioning young adults.

Studies that consistently reported the poorest validity coefficients used gifted samples (Livesay & Mealor, 1987a, 1987b; Phelps, 1989), or samples of young children, or both (Carvajal, Hardy, Smith, & Weaver, 1988; Clark, Wortman, Warnock, & Swerdlik, 1987). In the case of gifted children, this suggests that one of the tests used in the studies may not have had an adequate number of ceiling items. The opposite may be true for younger children, where the number of floor items may have been inadequate. An alternative explanation for both populations may be that the SB:FE and other intelligence tests are measuring different constructs for children who are younger or gifted or both.

It is interesting to note that most studies using abbreviated batteries of the SB:FE (e.g., Carvajal & Weyand, 1986; Hartwig et al., 1987; Hayden et al., 1988) found results similar to those using the complete SB:FE. This suggests that the abbreviated batteries provide as valid a measure of intellectual ability as the complete SB:FE.

When examining the studies in Table 2, there are several important methodological and psychometric issues to keep in

mind. Almost all of the studies suffer from a small sample size. Most use samples of less than 50 subjects, with some using as few as 18 (Smith et al., 1989). Restriction of range, specifically incidental selection, is a problem in the studies that used special populations in which selection was based in part on performance on a criterion (e.g., another intelligence test, an achievement test) that is correlated with the SB:FE. Other psychometric issues to consider include the item gradients, reliability differences, skill differences, and the representativeness of the norming samples of the tests used (Bracken, 1988).

Table 3 contains comparisons between the area scores of the SB:FE and the Verbal and Performance scales of the Wechsler tests and the Mental Processing scales of the K-ABC. It is more difficult to interpret the comparisons contained in Table 3, because the theoretical underpinnings and nature of the tasks composing the subscales within the different intelligence tests vary. Related to this point, it was previously established that the factor structure of the SB:FE may be different from that proposed by the test authors. This confuses the issue of comparison between the different factors on the SB:FE and those on other intelligence tests. Nevertheless, an attempt is made to discuss general trends suggested by Table 3.

Again, some arbitrary standards must be provided before examining Table 3. For instance, which K-ABC scales are most likely to correlate with which SB:FE area scores? One would expect the Short-Term Memory area score to correlate more highly with the Sequential Processing scale than with the Simultaneous Processing scale because of the similarity between the nature of the tasks that compose the Short-Term Memory and Sequential Processing scales. After this general statement, comparisons become more difficult. The only number-oriented task on the K-ABC, Number Recall, falls on the Sequential Processing scale. However, because of the visually oriented nature of the tasks on the Quantitative Reasoning factor, one might expect higher correlations between the Quantitative Reasoning area score and the Simultaneous Processing scale than between the Quantitative Reasoning area score and Sequential Processing scale. The Abstract/Visual Reasoning area score would seem more likely to be correlated with subtests such as Gestalt Closure, Matrix Analogies, and Spatial Memory—all members of the Simultaneous Processing scale; therefore, one would expect a higher correlation between these two scales than between Abstract/Visual Reasoning and Sequential Processing. Finally, the Verbal Reasoning area score would appear more likely to correlate with the Sequential Processing scale than it would with the Simultaneous Processing scale.

If these assumptions are accepted and used in examining Table 3, the support for these expected relationships is mixed. With the exception of Hayden et al. (1988), most studies have found that the correlations between the Short-Term Memory area score and the Sequential Processing scale are higher than those between the Short-Term Memory area score and the Simultaneous Processing scale. The comparisons between the Quantitative Reasoning area score and Sequential and Simultaneous Processing scales find correlations typically in the direction of the Simultaneous Processing scale, which is consistent with what was predicted. Mixed support is found when examining the relationship between the Abstract/Visual Reasoning area score and the K-ABC processing scales. When there is a difference, it is consistent with the direction predicted (i.e., the Abstract/Visual Reasoning area score is more highly correlated with the Simultaneous Processing scale). However, in several studies there was only a minimal difference in the correlations of the Abstract/Visual Reasoning area score and the Simultaneous and Sequential Processing scales. Finally, again with the exception of Hayden et al., the Verbal Reasoning area score tended to be more highly correlated with the Sequential Processing scale than with the Simultaneous Processing scale. Because the Hayden et al. sample consisted of gifted children, restriction of range may have contributed to the discrepant findings. In general, it appears that the relationship between the SB:FE area scores and the K-ABC processing scales is consistent with intuitive predictions of what the scales measure.

A similar approach can be used in examining the relationship between the SB:FE area scores and the Verbal and Performance scores of the Wechsler scales. One would expect the Verbal Reasoning area score to correlate more highly with the Verbal rather than with the Performance scale of the Wechsler scales. The reverse is true for the Abstract/Visual Reasoning area score: The Performance rather than the Verbal scale would be expected to be more highly correlated. The Quantitative Reasoning area score would be predicted to correlate more highly with the Verbal scale because the Arithmetic subtest is on this scale. Finally, the SB:FE Short-Term Memory area score should correlate more highly with the Verbal scale because the Digit Span and Arithmetic subtests are associated with short-term memory. When examining Table 3, it is apparent that these general predictions are supported in varying degrees. Again, these are intuitive predictions and do not provide empirical support for the factor structure of the SB:FE.

In sum, the results of the criterion-related studies involving the SB:FE seem to indicate that this instrument does measure intellectual functioning. It appears to be at least as good a measure of g as other existing measures of intelligence, especially for nonexceptional populations. The low coefficients obtained in studies using young children and gifted students lead one to question the use of the SB:FE with these populations or at least whether the test is measuring something other than the factors proposed by the authors. However, certain psychometric considerations (e.g., the publication dates and representativeness of the norming samples of the tests used) must be kept in mind when reviewing these results.

Other Measures of the SB:FE's Validity

The construct validity of the SB:FE also can be examined by determining its ability to differentiate previously identified groups of exceptional students who are expected to perform differently on the SB:FE. The intellectual functioning of different groups of students is expected to vary according to their classification (i.e., mentally retarded, gifted). If the SB:FE can discriminate between groups of exceptional youngsters, it would be consistent with this view of intelligence, and we could conclude that the SB:FE demonstrated construct validity.

The *Technical Manual* provides data on three groups of students—gifted, learning disabled, and mentally retarded—who were identified through procedures used by their respective school or clinic. The mean scores for the various areas and the

Table 3
Comparisons Between the Area Scores of the Stanford-Binet: Fourth Edition and the Scores on the Scales of the Kaufman Assessment Battery for Children (K-ABC), Wechsler Adult Intelligence Scale—Revised (WAIS-R), Wechsler Intelligence Scale for Children—Revised (WISC-R), and Wechsler Preschool and Primary Scale of Intelligence (WPPSI)

Measure	Sample	Verbal Reasoning	Abstract/Visual Reasoning	Quantitative Reasoning	Short-Term Memory
Technical Manual	Normal				
K-ABC SEQ		.77	.68	.73	.82
K-ABC SIM		.71	.77	.72	.73
Smith & Bauer (1989)	Normal				
K-ABC SEQ		.63*	.40*	.54*	.62*
K-ABC SIM		−.35*	.37*	.08	−.03
Technical Manual	Learning disabled				
K-ABC SEQ		.54	.28	.31	.55
K-ABC SIM		.33	.50	.47	.43
Smith et al. (1989)	Learning disabled				
K-ABC SEQ		.55*	.01	.54*	.70*
K-ABC SIM		.26	.64*	.60*	.22
Hayden et al. (1988)	Gifted				
K-ABC SEQ		.26	.56*	.35	.30
K-ABC SIM		.45*	.50*	.50*	.42
Krohn & Lamp (1989)	Head Start				
K-ABC SEQ		.76*	.66*	.74*	.78*
K-ABC SIM		.76*	.63*	.59*	.69*
Technical Manual	Normal				
WAIS-R VIQ		.86	.68	.85	.82
WAIS-R PIQ		.79	.81	.80	.65
Carvajal et al. (1987)	Normal				
WAIS-R VIQ		.58*	.53*	.77*	.49*
WAIS-R PIQ		.46*	.78*	.39*	.40*
Goh & Cordoni (1989)	College learning disabled				
WAIS-R VIQ		.65*	.43*	.63*	.57*
WAIS-R PIQ		.05	.63*	.49*	.25
Richardson (1988)	Educable mentally handicapped				
WAIS-R VIQ		.64*	.23	.46*	.32
WAIS-R PIQ		.24	.54*	.05	.25
Technical Manual	Mentally retarded				
WAIS-R VIQ		.59	.28	.65	.76
WAIS-R PIQ		.34	.61	.64	.56
Technical Manual	Normal				
WISC-R VIQ		.72	.68	.64	.64
WISC-R PIQ		.60	.67	.63	.63
Rothlisberg (1987)	Normal				
WISC-R VIQ		.70*	.43*	.17	.53*
WISC-R PIQ		.07	.55*	.40*	.41*
Swerdlik & Ryburn (1989)	Referred				
WISC-R VIQ		.89*	.50*	.60*	.67*
WISC-R PIQ		.61*	.56*	.66*	.57*
Technical Manual	Gifted				
WISC-R VIQ		.71	.21	.67	.44
WISC-R PIQ		.50	.43	.48	.63
Phelps (1989)	Gifted				
WISC-R VIQ		.36*	−.22	.10	.51*
WISC-R PIQ		−.08	.22	.00	.31
Technical Manual	Learning disabled				
WISC-R VIQ		.83	.68	.69	.69
WISC-R PIQ		.59	.72	.52	.55
Phelps et al. (1988)	Learning disabled				
WISC-R VIQ		.77*	.49*	.70*	.65*
WISC-R PIQ		.51*	.79*	.61*	.58*
Technical Manual	Mentally retarded				
WISC-R VIQ		.66	.32	.54	.52
WISC-R PIQ		.22	.55	.20	.38
Technical Manual	Normal				
WPPSI VIQ		.80	.46	.70	.71
WPPSI PIQ		.63	.56	.66	.59

Note. Technical Manual does not contain level of significance for analyses. SEQ = Sequential Processing; SIM = Simultaneous Processing; VIQ = Verbal IQ; PIQ = Performance IQ.
* Correlation significant at least $p < .05$.

SB:FE Test Composite for gifted students ranged from 1.2 to 1.5 standard deviations above the means of the standardization sample. The means for the learning disabled and mentally retarded students ranged from 0.7 to 1.0 and 2.4 to 2.8 standard deviations below those of the standardization sample, respectively (R. L. Thorndike et al., 1986b). These results suggest that the SB:FE can discriminate between different populations of exceptional children.

Further support for the SB:FE's ability to differentiate groups of exceptional children came in a study of normal, neurodevelopmentally impaired (e.g., learning disability, attention deficit disorder), and frank neurologically impaired (e.g., cerebral palsy, spina bifida) preschoolers (Hooper, Mayes, Swerdlik, & McNelis, 1990). The area scores and Test Composite score for the normal children were significantly higher than were those for the neurodevelopmentally and frank neurologically impaired preschoolers. These findings are consistent with those reported in the *Technical Manual* and suggest that the SB:FE can discriminate between normal and exceptional samples. However, Hooper et al. (1990) reported that no significant differences existed between the two impaired groups. These findings question the ability of the SB:FE to discriminate between the unique groups of exceptional youngsters included in this study. Whether one should expect differences on the SB:FE between these groups of youngsters who both suffer some sort of neurological impairment is unclear, and Hooper et al. do not suggest so.

Finally, it is expected that tests of intelligence be able to predict school achievement (Salvia & Ysseldyke, 1988). In other words, high scores on the SB:FE should correlate with high levels of school achievement. Indeed, this was the case in a study of 80 regular education third grade students (Powers, Church, & Treloar, 1989). The SB:FE Test Composite score was the single best predictor of all the areas of achievement measured on the Woodcock-Johnson Tests of Achievement (i.e., reading, mathematics, and written language). Similarly, the Test Composite score was significantly correlated with Metropolitan Readiness Test scores for preschoolers of low socioeconomic status (Krohn & Lamp, 1990). Finally, the *Technical Manual* (R. L. Thorndike et al., 1986b) contains information regarding the SB:FE and the Achievement scale from the K-ABC. The correlations between the SB:FE area scores and the Achievement scale ranged from .45 to .72, whereas the correlation between the Test Composite and Achievement scale was .74. The magnitude of these correlations suggests that performance on the SB:FE is related to scores on the K-ABC measure of achievement. It is noteworthy that Powers et al. found that the SB:FE area and Test Composite scores were better predictors of academic achievement than were Sattler's (1988) factor scores.

The fact that the SB:FE was able to differentiate between different groups of exceptional children suggests that the instrument does demonstrate construct validity. The high correlations between the SB:FE Test Composite and scores on achievement tests also supports the construct validity of this measure.

Summary and Conclusions

After reviewing the validity studies available regarding the SB:FE, several conclusions can be drawn. The SB:FE appears to be at least as good a measure of g as other currently available tests of intellectual ability. In addition to measuring g, the SB:FE measures other factors. Unfortunately, there is mixed support as to whether the factors proposed by the test authors actually are measured by the SB:FE. In general, there is more evidence that the SB:FE is not a four-factor test but rather a two-factor test (i.e., Verbal, Nonverbal) for children through the age of 6 years (Keith et al., 1988; Kline, 1989; Reynolds et al., 1988; Sattler, 1988; R. M. Thorndike, 1990) and a three-factor test (i.e., Verbal, Nonverbal, Memory) for individuals beyond age 6 years (Kline, 1989; Sattler, 1988; R. M. Thorndike, 1990). Even in situations in which the factors resemble those hypothesized by R. L. Thorndike et al. (1986a, 1986b), there is not perfect agreement. In other words, some of the subtests load as well or better on factors other than those hypothesized (e.g., Memory for Sentences). In addition, the Quantitative Reasoning factor does not appear to consistently load on the Crystallized Abilities factor, which constitutes the second level of the hierarchical structure (Keith et al., 1988).

The lowest concurrent criterion-related validity coefficients were obtained for gifted students and younger children. These results may reflect psychometric phenomena such as restricted range or the effects of newer normative data (Bracken, 1988). With regard to gifted students, lower concurrent validity coefficients, particularly those between the SB:L-M and SB:FE, may reflect the heavier emphasis on verbal content in the SB:L-M. The broader nature of subtests on the SB:FE may make it more difficult for verbally oriented youngsters to score in a range that would qualify them as "gifted" in existing systems.

Other construct validity studies suggest that the SB:FE can distinguish between groups of youngsters with differing intellectual abilities (e.g., mentally handicapped, gifted, neurologically impaired). In addition, there is a positive correlation between scores on the SB:FE and measures of achievement.

On the basis of the information provided in this article, it seems appropriate to use the SB:FE as a measure of intellectual functioning in a number of situations. In general, the SB:FE seems to be as valid as any existing measure of intellectual functioning. Therefore, it is useful in most instances in which a measure of g is important. The Test Composite score derived from the SB:FE could be used as a measure of ability in the ability–achievement discrepancy model in determining whether a child is experiencing a learning disability. That the factor structure of the SB:FE may not be consistent with that proposed by the test authors suggests that an alternative interpretive system such as that provided by Sattler (1988) may be important to consider when using the results of the SB:FE beyond the Test Composite score.

The SB:FE would appear to be useful in cases in which concerns exist regarding short-term memory. Regardless of the interpretive system used (i.e., R. L. Thorndike et al., 1986a; Sattler, 1988), a Memory factor emerges after age 6 years. Other popular tests of intellectual functioning (e.g., WISC-R, K-ABC) were not so designed and therefore do not profess to directly measure memory.

Despite the low validity coefficients between the full scale scores of other measures of intelligence (e.g., K-ABC, SB:L-M, WISC-R) and the SB:FE, the SB:FE may be the instrument of choice in assessing children referred for gifted programs. The

SB:FE measures a broad range of abilities and provides more challenging items and a higher ceiling for bright youngsters in the early adolescent years because of the age range covered by the test. Other measures of intelligence that cover a more limited age range typically do not provide enough difficult items for gifted youngsters near the upper age extreme of the test.

Although the SB:FE can be used with children as young as 2 years of age, this does not imply that it is necessarily the instrument of choice for use with low-functioning individuals. The test lacks floor items low enough for younger children, making it difficult to diagnose mental retardation, especially for 2- and 3-year-olds (Sattler, 1988). On the other hand, the wide variety of tasks and broad age range covered by the SB:FE appear to provide enough floor items to assess the abilities of young adults with moderate mental retardation.

The SB:FE also may be the instrument of choice when an examiner has a limited amount of time to work with a child because of the defined abbreviated batteries. As was seen in a previous section, the validity coefficients between the full scale scores of other measures of intelligence (e.g., K-ABC, SB:L-M, WISC-R) and abbreviated batteries of the SB:FE were similar to those of the complete SB:FE. Although short forms of other measures of intelligence exist, they have been generated subsequent to the publication of the test rather than addressed as part of test development. As a result, there is confusion about when to use one short form over another. R. L. Thorndike et al. (1986a) have provided specific guidelines for SB:FE abbreviated batteries.

The SB:FE may be the instrument of choice when the user is interested in measuring an individual's performance over time. The broad age range covered by the SB:FE allows more continuous measurement with one instrument than is available from other commonly used intelligence tests, for instance, with the individual Wechsler scales. However, although the same area scores are obtained across levels using the R. L. Thorndike et al. (1986a) scoring system, the SB:FE is not truly a continuous measure across ages. It is important to recognize that different subtests compose the area scores at different age levels. In other words, not all subtests are given to all examinees, meaning individuals are administered a different test battery depending on their age. In addition, the maximum and minimum scores that can be obtained on the SB:FE fluctuate across ages (Sattler, 1988). As Sattler pointed out, perfect performance at ages 12 and 18 years can result in almost a standard deviation drop in Test Composite scores, although the individual performed virtually identically from one time to the other. Continuous measurement is useful, but it should not lull one into a false sense of security in interpreting test results, especially in the case of the SB:FE.

In closing, it is difficult to make a definitive statement about the use of any test of intellectual functioning. As is the case with most other intelligence tests, the SB:FE provides a good measure of general ability. That may be all that can be expected of any traditional measure of intelligence. The support for the hypothesized factors of the SB:FE is mixed. The most prudent approach to interpretation beyond the Test Composite score seems to be that provided by Sattler (1988), who views the SB:FE as a two-factor test for ages 2 through 6 years and a three-factor test beyond age 6 years. Sattler's interpretive approach seems to be the one best supported by research, and it provides helpful information when using the SB:FE. There will be some instances in which the SB:FE provides useful information and others in which another test will provide more meaningful data. Rarely does one test do everything one would like it to do. The determining factor in deciding to use the SB:FE should be the type of information one is interested in and whether that information can be obtained using the SB:FE.

References

Aiken, L. R. (1987). *Assessment of intellectual functioning.* Boston: Allyn & Bacon.

Boyle, G. J. (1989). Confirmation of the structural dimensionality of the Stanford-Binet Intelligence Scale (Fourth Edition). *Personality and Individual Differences, 10,* 709–715.

Bracken, B. A. (1988). Ten psychometric reasons why similar tests produce dissimilar results. *Journal of School Psychology, 26,* 155–166.

Carvajal, H., Gerber, J., Hewes, P., & Weaver, K. A. (1987). Correlations between scores on Stanford-Binet IV and Wechsler Adult Intelligence Scale—Revised. *Psychological Reports, 61,* 83–86.

Carvajal, H., Hardy, K., Smith, K. L., & Weaver, K. A. (1988). Relationships between scores on Stanford-Binet IV and Wechsler Preschool and Primary Scale of Intelligence. *Psychology in the Schools, 25,* 129–131.

Carvajal, H., & Weyand, K. (1986). Relationships between scores on Stanford-Binet IV and Wechsler Intelligence Scale for Children—Revised. *Psychological Reports, 59,* 963–966.

Cattell, R. B. (1963). Theory for fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,* 1–22.

Chattin, S. H., & Bracken, B. A. (1989). School psychologists' evaluation of the K-ABC, McCarthy Scales, Stanford-Binet IV, and WISC-R. *Journal of Psychoeducational Assessment, 7,* 112–130.

Clark, R. D., Wortman, S., Warnock, S., & Swerdlik, M. (1987). A correlational study of Form L-M and the 4th edition of the Stanford-Binet with 3- to 6-year olds. *Diagnostique, 12,* 118–120.

Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology, 55,* 584–594.

Davis, H. E. (1989, July/August). TESTimonial: The new Stanford-Binet: Unfriendly to the user. *School Psychology in Illinois,* 13.

Freides, D. (1972). Stanford-Binet Intelligence Scale. Third Edition. In O.K. Buros (Ed.), *The seventh mental measurements yearbook* (pp. 772–773). Highland Park, NJ: Gryphon Press.

Glutting, J. J. (1989). Introduction to the structure and application of the Stanford-Binet Intelligence Scale—Fourth Edition. *Journal of School Psychology, 27,* 69–80.

Goh, D. S., & Cordoni, B. (1989, March). *Comparison of the Stanford-Binet Fourth Edition with WAIS-R in college learning disabled students.* Paper presented at the Annual Convention of the National Association of School Psychologists, Boston.

Hanson, D. P. (1989, April). *The new Stanford-Binet: Practical considerations.* Paper presented at the Annual Convention of the National Association of School Psychologists. San Francisco.

Hartwig, S. S., Sapp, G. L., & Clayton, G. A. (1987). Comparison of the Stanford-Binet Intelligence Scale: Form L-M and the Stanford-Binet Intelligence Scale: Fourth Edition. *Psychological Reports, 60,* 1215–1218.

Hayden, D. C., Furlong, M. J., & Linnemeyer, S. (1988). A comparison of the Kaufman Assessment Battery for Children and the Stanford-Binet IV for the assessment of gifted children. *Psychology in the Schools, 25,* 239–243.

Heath, W. M. (1988, Fall). The use and acceptance of the Stanford-Binet: Fourth Edition. *New York State School Psychology Newsletter.*

Hooper, S. R., Mayes, S. D., Swerdlik, M. E., & McNelis, M. M. (1990, April). *A comparison between normal, neurodevelopmentally impaired, and frank neurologically impaired preschool children on the Stanford-Binet Intelligence Scale: Fourth Edition.* Paper presented at the Annual Convention of the National Association of School Psychologists, San Francisco.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology, 57,* 253–270.

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

Kaufman, A. S., & Kaufman, N. L. (1983). *K-ABC: Kaufman Assessment Battery for Children.* Circle Pines, MN: American Guidance Service.

Keith, T. Z., Cool, V. A., Novak, C. G., White, L. J., & Pottebaum, S. M. (1988). Confirmatory factor analysis of the Stanford-Binet Fourth Edition: Testing the theory-test match. *Journal of School Psychology, 26,* 253–274.

Kline, R. B. (1989). Is the fourth edition Stanford-Binet a four-factor test? Confirmatory factor analyses of alternative models for ages 2 through 23. *Journal of Psychoeducational Assessment, 7,* 4–13.

Krohn, E. J., & Lamp, R. E. (1989). Concurrent validity of the Stanford-Binet Fourth Edition and K-ABC for Head Start children. *Journal of School Psychology, 27,* 59–67.

Krohn, E. J., & Lamp, R. E. (1990, August). *Predictive validity of K-ABC, Binet-LM, and Binet-4 with preschoolers.* Paper presented at the 98th Annual Convention of the American Psychological Association, Boston.

Livesay, K. K., & Mealor, D. J. (1987a, March). *A comparative analysis of the Stanford-Binet: Fourth Edition and WISC-R for gifted referrals.* Paper presented at the Annual Convention of the National Association of School Psychologists, New Orleans.

Livesay, K. K., & Mealor, D. J. (1987b, March). *A comparative of the Stanford-Binet: Fourth Edition and Form L-M for gifted referrals.* Paper presented at the Annual Convention of the National Association of School Psychologists, New Orleans, LA.

Long, J. S. (1983). *Confirmatory factor analysis.* Beverly Hills, CA: Sage.

Lubin, B., Larsen, R. M., & Matarazzo, J. D. (1984). Patterns of psychological test usage in the United States: 1935–1982. *American Psychologist, 39,* 451–454.

Lukens, J. (1988). Comparison of the fourth edition and L-M edition of the Stanford-Binet used with mentally retarded persons. *Journal of School Psychology, 26,* 87–89.

McCall, V. W., Yates, B., Hendricks, S., Turner, K., & McNabb, B. (1989). Comparison between the Stanford-Binet: L-M and the Stanford-Binet: Fourth Edition with a group of gifted children. *Contemporary Educational Psychology, 14,* 93–96.

Ownby, R. L., & Carmin, C. N. (1988). Confirmatory factor analyses of the Stanford-Binet Intelligence Scale, Fourth Edition. *Journal of Psychoeducational Assessment, 6,* 331–340.

Phelps, L. (1989). Comparison of scores for intellectually gifted students on the WISC-R and the fourth edition of the Stanford-Binet. *Psychology in the Schools, 26,* 125–129.

Phelps, L., Bell, M. C., & Scott, M. J. (1988). Correlations between the Stanford-Binet: Fourth Edition and the WISC-R with a learning disabled population. *Psychology in the Schools, 25,* 380–382.

Powers, A. D., Church, J. K., & Treloar, J. (1989, March). *The fourth edition of the Stanford-Binet Intelligence Scale and the Woodcock-Johnson Tests of Achievement: A criterion validity study.* Paper presented at the Annual Meeting of the National Association of School Psychologists, Boston.

Reynolds, C. R. (1987). Playing IQ roulette with the Stanford-Binet, 4th Edition. *Measurement and Evaluation in Counseling and Development, 3,* 139–141.

Reynolds, C. R., Kamphaus, R. W., & Rosenthal, B. L. (1988). Factor analysis of the Stanford-Binet Fourth Edition for ages 2 years through 23 years. *Measurement and Evaluation in Counseling and Development, 21,* 52–63.

Richardson, J. C. (1988). *Coefficients of correlations of IQ's on the WAIS-R with standard age scores on the Stanford-Binet, 4th Edition for previously identified mentally handicapped adolescents.* Paper presented at the Annual Meeting of the National Association of School Psychologists, Chicago.

Rothlisberg, B. A. (1987). Comparing the Stanford-Binet, Fourth Edition to the WISC-R: A concurrent validity study. *Journal of School Psychology, 25,* 193–196.

Rothlisberg, B. A., McIntosh, D. E., & Dodge, J. K. (1990, April). *Comparability of the Stanford-Binet IV to the K-ABC in a referred sample.* Paper presented at the Annual Meeting of the National Association of School Psychologists, San Francisco.

Salvia, J., & Ysseldyke, J. E. (1988). *Assessment in special and remedial education. (4th ed.).* Boston: Houghton Mifflin.

Sattler, J. M. (1988). *Assessment of children. (3rd ed.).* San Diego, CA: Author.

Smith, D. K., & Bauer, J. J. (1989, March). *Relationship of the K-ABC and S-B:FE in a preschool sample.* Paper presented at the Annual Meeting of the National Association of School Psychologists, Boston.

Smith, D. K., St. Martin, M. E., & Lyon, M. A. (1989). A validity study of the Stanford-Binet: Fourth Edition with students with learning disabilities. *Journal of Learning Disabilities, 22,* 260–261.

Swerdlik, M. E., & Ryburn, M. (1989, April). *A construct validity study of the Stanford-Binet Fourth Edition and Wechsler Intelligence Scale for Children—Revised.* Paper presented at the Annual Convention of the National Association of School Psychologists, Boston.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986a). *The Stanford-Binet Intelligence Scale: Fourth Edition. Guide for administering and scoring. (2nd printing).* Chicago: Riverside.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986b). *The Stanford-Binet Intelligence Scale: Fourth Edition. technical manual.* Chicago: Riverside.

Thorndike, R. M. (1990). Would the real factors of the Stanford-Binet Fourth Edition please come forward? *Journal of Psychoeducational Assessment, 8,* 412–435.

Woodcock, R. W., & Johnson, M. B. (1977). *Examiner's manual for the Woodcock-Johnson Psycho-Educational Battery, Part One: Tests of Cognitive Ability.* Allen, TX: DLM.

Received January 4, 1991
Revision received May 9, 1991
Accepted May 9, 1991 ■