(2003). In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment: Intelligence, aptitude, and achievement (2nd ed.) (pp. 217-242). New York: Guilford.

10

Stanford–Binet Intelligence Scale: Fourth Edition (SB4): Evaluating the Empirical Bases for Interpretations

ERIC A. YOUNGSTROM
JOSEPH J. GLUTTING
MARLEY W. WATKINS

The Stanford–Binet Intelligence Scale: Fourth Edition (SB4; Thorndike, Hagen, & Sattler, 1986b) is the most recent edition in a line of instruments going back almost a century (viz., Binet & Simon, 1905). The 1986 revision revitalized the Stanford–Binet by both maintaining links with previous editions of the scale and simultaneously incorporating more recent developments found in other popular tests of intelligence. The SB4 retains as much item content as possible from the Stanford–Binet Intelligence Scale: Form L-M (SB-LM; Thorndike, 1973). SB4 also respects tradition by covering approximately the same age range as SB-LM (ages 2–23); it incorporates familiar basal and ceiling levels during testing; and it provides an overall score that appraises general cognitive functioning. As this chapter is being written, the fifth edition of the Stanford–Binet (SB5) is beginning item tryouts in preparation for standardization. The plan for this newest edition also shares a commitment both to the Stanford–Binet tradition and to incorporating current theories about psychometrics (e.g., item response theory) and the structure of intelligence.

Despite these similarities, these revisions are substantially different from their predecessors. The SB4 eliminated the traditional age scale format. In its place are 15 subtests whose age-corrected scaled scores make it possible to interpret profile elevations and profile depressions. Four "area" scores, derived from theoretically based subtest groupings, are also new. These reformulations add to interpretative possibilities, and they attempt to broaden the coverage of cognitive ability over that offered by SB-LM. SB4 permits calculation of the Composite (overall IQ) for performances based on specific "abbreviated batteries," as well as for *any* combination of subtests psychologists wish to regroup—promoting flexibility in administration and interpretation.

This chapter familiarizes readers with the structure and content of SB4. It also evaluates selected aspects of the test's psychometric and technical properties. In addition, we hope to sensitize psychologists to factors pertinent to the administration of SB4 and to the interpretation of its test scores. The chapter aims to present a balanced treat-

ment of strengths and limitations. The highest professional standards were applied throughout the development of SB4. Prior to publication, the authors and publisher dedicated over 8 years to development and 2 years to extensive data analyses of the final product. Thus, it is no surprise that we identify unique and praiseworthy features. Similarly, no test is without faults, and this thought should place the potential shortcomings of SB4 in context.

The first section of this chapter outlines the theoretical model underlying SB4. We turn then to a general description of the structure of SB4 and issues related to its test materials, administration, and scaling. Thereafter, we discuss the strengths and weaknesses associated with SB4's standardization, its reliability and validity, and factors related to the interpretation of its test scores. Finally, we take a look at the development of SB5.

THEORETICAL FOUNDATION

Perhaps the most fundamental change incorporated into SB4 is the expansion of its theoretical model. Figure 10.1 shows that SB4 has three levels, which serve both traditional and new Binet functions. At the apex is the Composite, or estimate of general ability, traditionally associated with Binet scales. The second level is new to SB4. It proposes three group factors: Crystallized Abilities, Fluid-Analytic Abilities, and Short-Term Memory. The first two dimensions originate from the Cattell–Horn theory of intelligence (Cattell, 1940; Horn, 1968; Horn & Cattell, 1966). These are shaded in the figure, because the published interpretive system for the SB4 does not emphasize the calculation of observed scores corresponding with these factors. The additional component, Short-Term Memory, is not contained in the Cattell–

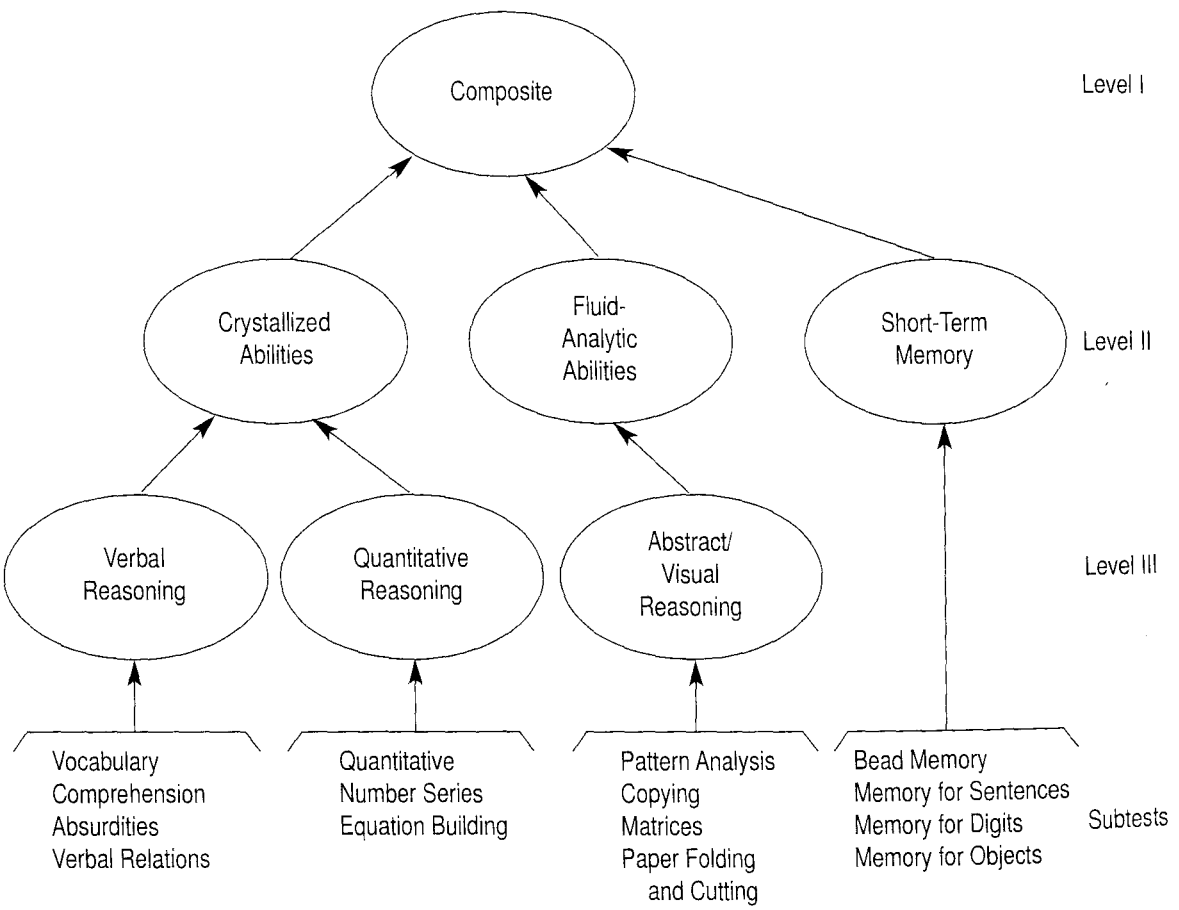

FIGURE 10.1. Theoretical model for SB4.

Horn theory. Its inclusion reflects the way in which psychologists used previous editions of the Binet (Thorndike, Hagen, & Sattler, 1986a); to some extent, it also reflects factor-analytic work with other intelligence tests, suggesting that short-term memory is related to long-term memory and to more complex learning and problem solving (Thorndike, Hagen, & Sattler, 1986c).

The third level illustrates another difference between the SB4 and earlier editions of the scale. Here, factors are identified in terms of three facets of reasoning: Verbal Reasoning, Quantitative Reasoning, and Abstract/Visual Reasoning. These components resemble the third level of Vernon's (1950) hierarchical model of intelligence, wherein well-known Verbal–Educational and Practical–Mechanical factors are subdivided to obtain even more homogeneous estimates of ability. Vernon, for example, splits the Verbal–Educational factor into the scholastic content of verbal fluency, numerical operations, and so on. SB4 follows this orientation by incorporating dimensions for the assessment of Verbal Reasoning and Quantitative Reasoning. Similarly, SB4's Abstract/Visual Reasoning dimension parallels the Practical–Mechanical component of the Vernon model.

The three group factors at the third level (Verbal Reasoning, Quantitative Reasoning, Abstract/Visual Reasoning), plus the Short-Term Memory factor at the second level, form the four "area" scores derived by SB4. The Abstract/Visual Reasoning score at the third level corresponds to the Fluid-Analytic Abilities dimension at the second level. No area score is readily available for the third dimension at the second level, Crystallized Abilities; nevertheless, scores for this broad-band factor can be estimated by collapsing results across the remaining two of the four areas (Verbal Reasoning and Quantitative Reasoning).

Thus the SB4 model is an eclectic unification of multiple theories of intelligence. Such synthesis is not unique to SB4. The Kaufman Assessment Battery for Children (K-ABC; Kaufman & Kaufman, 1983) also accounts for test performance through interrelationships among theories (i.e., the Luria–Das and Cattell–Horn theories of ability). In addition, both tests share the desirable quality of using explicit theoretical frameworks as guides for item development and for the alignment of subtests within modeled hierarchies.

TEST STRUCTURE

Subtest Names and Content

SB4's subtests retain some reliable variance that is distinct from the score variation captured by area scores or the Composite. Because of this specificity within each subtest, the developers described the "unique abilities" evaluated by SB4's subtests. Profile analysis is a popular method for explicating an examinee's strengths and weaknesses on these abilities (see Delaney & Hopkins, 1987; Naglieri, 1988a, 1988b; Rosenthal & Kamphaus, 1988; Sattler, 1992; Spruill, 1988). Therefore, inasmuch as SB4 supports comparisons among subtest scores, it is worthwhile to understand the identity and composition of these measures.

Descriptions of the 15 SB4 subtests are provided in Table 10.1, organized according to the theoretical area each occupies in the scale.

Content Similarity with Other IQ Tests

SB4 items appear representative of the item content found in intelligence tests (see Jensen, 1980, for a detailed analysis of item types common among IQ tests). Visual inspection reveals that six SB4 subtests share core content with the Wechsler Intelligence Scale for Children—Third Edition (WISC-III; Wechsler, 1991). For example, both SB4 Vocabulary and WISC-III Vocabulary assess word knowledge. SB4 Comprehension and WISC-III Comprehension measure breadth of knowledge of social and interpersonal situations, and the visual-perceptual abilities evaluated by SB4 Pattern Analysis generally apply to WISC-III Block Design. Likewise, there are marked similarities between the SB4 Quantitative subtest and WISC-III Arithmetic, between SB4 Memory for Digits and WISC-III Digit Span, and between SB4 Verbal Relations and WISC-III Similarities. Resemblances in subtest content are also ap-

TABLE 10.1. SB4 Subtests: Age Range, Median Reliability, and Content

Area/subtest	Ages	Reliability	Content
Verbal Reasoning			
Vocabulary	2–23	.87	Examinees supply word definitions. The first 15 items tap receptive word knowledge (examinees name pictured objects), and items 16 through 46 are presented both orally and in writing vocabulary.
Comprehension	2–23	.89	Items 1 through 6 require the receptive identification of body parts. Items 7 through 42 elicit verbal responses associated with practical problem solving and social information.
Absurdities	2–14	.87	This subtest presents situations that are essentially false or contrary to common sense. Examinees point to the inaccurate picture among three alternatives (items 1 through 4), or they verbalize the absurdity in a single picture (items 5 through 32).
Verbal Relations	12–23	.91	Examinees state how three words, out of a four-word set, are similar. The fourth word in each item is always different from the three words preceding it.
Quantitative Reasoning			
Quantitative	2–23	.88	Examinees are required to count, add, seriate, or complete other numerical operations (e.g., count the number of blocks pictured; how many 12″ by 12″ tiles would be needed to cover a floor that is 7 feet by 9 feet?).
Number Series	7–23	.90	A row of four or more numbers is presented, and the task is to identify the principle underlying a series of four or more numbers and to apply that principle to generate the next two numbers in the series (e.g., 1, 3, 7, 15, __, __).
Equation Building	12–23	.91	Examinees resequence numerals and mathematical signs into a correct solution (e.g., 15, 12, 2, 25, =, +, −).
Abstract/Visual Reasoning			
Pattern Analysis	2–23	.92	Items 1 through 6 require examinees to complete formboards. Items 7 through 42 involve the replication of visual patterns through block manipulations.
Copying	2–13	.87	Examinees either reproduce block models (items 1 through 12) or draw geometric designs, such as lines, rectangles, and arcs, that are shown on cards (items 13 through 28).
Matrices	7–23	.90	Each item presents a matrix of figures in which one element is missing. The task is to identify the correct element among multiple-choice alternatives.
Paper Folding and Cutting	12–23	.94	Figures are presented in which a piece of paper has been folded and cut. Examinees chose among alternatives that show how the paper might look if it were unfolded.
Short-Term Memory			
Bead Memory	2–23	.87	Examinees recall the identity of one or two beads exposed briefly (items 1 through 10), or they reproduce bead models in a precise sequence (items 11 through 42).
Memory for Sentences	2–23	.89	Examinees are required to repeat each word in a sentence in the exact order of presentation.
Memory for Digits	7–23	.83	Examinees repeat digits either in the sequence they are presented, or in reverse order.
Memory for Objects	7–23	.73	Pictures of objects are viewed briefly. Examinees then identify the objects in correct order from a larger array.

parent between SB4 and the K-ABC. The four most striking parallels occur between (1) SB4 Pattern Analysis and K-ABC Triangles, (2) SB4 Matrices and K-ABC Matrix Analogies, (3) SB4 Memory for Digits and K-ABC Number Recall, and (4) SB4 Memory for Objects and K-ABC Word Order. These comparisons suggest that there exists a core set of subtests (generally including those with the highest *g* saturation) that are shared across commonly used measures of ability.

MATERIALS

Three manuals accompany SB4: the *Guide for Administering and Scoring* (Thorndike et al., 1986a), the *Technical Manual* (Thorndike et al., 1986c), and the supplementary *Examiner's Handbook* (Delaney & Hopkins, 1987). All three manuals are well written and informative. Chapters pertinent to test administration are especially well organized in the *Examiner's Handbook*. Psychologists new to SB4 are encouraged to read these sections of the handbook prior to reviewing the *Guide for Administering and Scoring*.

SB4 materials are attractive, well packaged, and suitable to the age groups for which they are applied. The Bead Memory subtest is a noteworthy exception. Directions for Bead Memory caution psychologists to **"BE SURE THAT EXAMINEES DO NOT PLAY WITH THE BEADS. THERE IS A DANGER THAT YOUNG EXAMINEES MAY TRY TO PUT THE BEADS IN THEIR MOUTHS"** (Thorndike et al., 1986b, p. 23; boldface capitals in original). This caution is insufficient for the danger presented. Two of the four bead types fit easily in a "choke tube"—an apparatus used to determine whether objects are sufficiently small that young children will gag or suffocate on them. Psychologists, therefore, should *never* allow young children to play with these objects.[1]

Publishers are increasingly adding color to test stimuli. Rich colors enhance the attractiveness of test stimuli, and they have the positive effect of making test materials more child-oriented (Husband & Hayden, 1996). Color helps to maintain children's interest during testing, and it augments the probability of obtaining valid test scores. However, sometimes color is not equally salient or perceptually unambiguous to examinees. Such a situation can arise when persons with color-blindness are being assessed. For these individuals, color represents an additional source of score variance that can reduce test validity. They are most likely to experience difficulty when confronted by the following color combinations: red–brown, green–orange, red–grey, blue–purple, and red–green (Coren, Ward, & Enns, 1999).

Two examples are presented where color may alter SB4 item difficulties. Item 1 of the Vocabulary subtest shows a red car on a brown background. This color combination makes it more difficult for some individuals with color-blindness to distinguish the important foreground stimulus (the car) from its background. Another example can be seen in the formboard items in Pattern Analysis. The red puzzle pieces and green background make the formboard more difficult for examinees with red–green color blindness. (See also Husband & Hayden, 1996, for an investigation of the effects of varying stimulus color on several SB4 subtests.) Fortunately, the problems associated with color stimuli can be corrected by simply *not* pairing these colors within test items. By adopting such changes, test publishers will be able to continue offering the benefits of color stimuli and simultaneously reduce the visual discrimination problems of examinees with color-blindness.

ADMINISTRATION

SB4 uses "adaptive testing" to economize on administration time. This format offers the added benefit of decreasing frustration, because examinees are exposed only to those test items most appropriate to their ability level. The Vocabulary subtest serves as a "routing" measure at the beginning of each assessment. Performance on the Vocabulary subtest, in conjunction with an examinee's chronological age, is used to determine the appropriate entry level for succeeding subtests. Entry levels are arranged hierarchically by item pairs (labeled "A" through "Q" on the test protocol). Basal and ceiling rules are then applied within subtests. A basal level is

established when all items are passed at two consecutive levels. A ceiling is reached, and testing advances to the next subtest, when three failures (out of four possible) take place across adjacent levels. There is some concern that the entry levels may be too high for youths and adults with mental retardation (Sattler, 1992; Spruill, 1991). This routing system also can be confusing to examiners unfamiliar with the SB4 testing format (Vernon, 1987; Wersh & Thomas, 1990). Supervisors and instructors should make certain that trainees are comfortable navigating the routing system (Choi & Proctor, 1994), and trainees should be vigilant for possible difficulty at subtest entry points when testing children with suspected cognitive deficits.

SB4 deserves credit for its efficient testing format and for directions that are readable and straightforward. In contrast to SB-LM, SB4 administration is simpler due to such features as incorporating most of the directions, stimuli, and scoring criteria within the easel kits. The use of sample items helps familiarize examinees with directions and item formats prior to actual testing. In addition, SB4 is a "power" test (as opposed to a "speeded" test; Anastasi & Urbina, 1997). Pattern Analysis is the only subtest requiring mandatory time limits. Doing away with the need for accurate timekeeping coincidentally makes SB4's administration more convenient.

Administration times appear reasonable. The *Technical Manual* (Thorndike et al., 1986c) does not offer administration times by age level. Delaney and Hopkins (1987) provide administration times by entry level (A through M or higher), and we used this information to approximate testing times by age. Based on these estimates, testing would take between 30 and 40 minutes for preschool-age children; 60 minutes for children between the ages of 6 and 11; and between 70 and 90 minutes for those at higher age levels. These values may underestimate actual testing times. Sattler (1992) reports that the full battery is much too long to complete in most circumstances, and he indicates that it may take 2 hours to administer the entire test to an adolescent. The length of time required for the full battery has spurred the development of a plethora of short forms, which are discussed below.

One final area of concern is the developmental appropriateness of an instrument for use with young children. Preschoolers vary in their knowledge of basic concepts (e.g., "top," "behind," "same as"). As a result, basic concepts in test directions may hinder preschool children's understanding of what is expected of them. Kaufman (1978) examined this issue by comparing the number of basic concepts in the Boehm Test of Basic Concepts (BTBC; Boehm, 1971) to those found in the directions for several preschool-level ability tests, including the following: SB-LM; the McCarthy Scales of Children's Abilities (MSCA; McCarthy, 1972); and the Wechsler Preschool and Primary Scale of Intelligence (WPPSI; Wechsler, 1967). Results revealed that scores from SB-LM (5 basic concepts) were less susceptible to this influence than scores from the MSCA (7 basic concepts) or WPPSI (14 basic concepts).

We compared directions in SB4 to basic concepts in the BTBC.[2] In particular, directions were analyzed for the eight SB4 subtests routinely administered to preschoolers. Our findings show that SB4 assumes young children know eight BTBC basic concepts. Although this represents an increase over the number found for SB-LM, it compared favorably to the number of basic concepts in the MSCA, and it is fewer than that found for the WPPSI. Thus SB4 directions are at least as likely to be understood by preschoolers as those contained in other IQ tests.

SCALING

Raw SB4 scores are converted to standard age scores (SASs). SASs for the four areas and the Composite are synonymous with deviation IQs ($M = 100$, $SD = 16$, consistent with Binet tradition). Subtest SASs are normalized standard scores with $M = 50$ and $SD = 8$. This metric is highly unusual. We find no compelling reasoning for this choice, and we share Cronbach's (1987) criticism of SB4 that there is no advantage for choosing these units over conventional T-scores.

Percentile ranks are available for subtests, area scores, and the Composite. Although SB4 is no longer an age scale, age equiva-

lents are supplied for the 15 subtests. Moreover, a conversion table is produced for professionals who wish to interpret area scores and the Composite in a metric identical to the Wechsler series (M = 100, SD = 15).

A historical advantage of Binet scales has been an extended floor for detecting moderate to severe mental retardation. Psychologists will be no doubt disappointed that this benefit is generally unavailable for young children on SB4 (Bradley-Johnson, 2001; Grunau, Whitfield, & Petrie, 2000; McCallum & Whitaker, 2000; Saylor, Boyce, Peagler, & Callahan, 2000). Table 10.2 presents minimum overall ability scores attainable for preschoolers on SB-LM, SB4, the WPPSI, and the K-ABC. Column 2 indicates that SB-LM was fully capable of diagnosing mild intellectual retardation by age 3, and moderate retardation by age 3 years, 6 months. In contrast, column 3 reveals that for all practical purposes, SB4's Composite is unable to diagnose mild intellectual deficits prior to age 4, and it shows no capacity for detecting moderate retardation until age 5.

Tests such as the WPPSI, WPPSI-R, and the K-ABC have been criticized for being insensitive to preschoolers who perform at the lower end of the ability continuum (Bracken, 1985; Olridge & Allison, 1968; Sattler, 1992). Column 5 in Table 10.2 shows that SB4 is somewhat more precise in this regard than the K-ABC. However, column 4 also reveals that SB4 is no more sensitive than the WPPSI-R. These comparisons, combined with existing views on the limitations of the WPPSI-R and K-ABC, lead to the conclusion that SB4 provides an insufficient floor for testing young children suspected to perform at lower levels of ability (Flanagan & Alfonso, 1995). These findings are disappointing, since SB-LM was the only IQ test capable of diagnosing mental retardation with preschoolers between the ages of 2 years, 6 months (the upper age range of the Bayley Scales) and 4 years, 0 months.

Problems are compounded for younger preschoolers by the fact that area scores evidence even higher floors than the Composite. For example, the lowest SAS for Quantitative Reasoning between the ages of 2 years, 0 months and 4 years, 6 months is 72. This score is above the range for mental retardation, and the median lowest attainable SAS is 87 for children between these ages. With this instrument, it is impossible for younger preschoolers to show deficient or abnormal functioning in Quantitative Reasoning. Even more disturbing, the truncated floor makes it more probable that an artifactual "pattern" of strength in Quantitative Reasoning will emerge for any such preschooler whose Composite is in the gender range. Floor limitations dissipate by the age of kindergarten entry. SB4's Composite is able to identify both mild and moderate intellectual retardation at the age of 5 years, 0 months. Similarly, shortcomings noted for the Quantitative Reasoning area are resolved essentially by the age of 4 years, 6 months (cf. Grunau et al., 2000).

Table 10.3 illustrates SB4's facility to detect functioning at the upper extreme. By in-

TABLE 10.2. Preschoolers' Minimum Overall Ability Scores on SB-LM, SB4, the WPPSI-R, and the K-ABC

Age in years and months	SB4[a]	SB-LM	WPPSI-R	K-ABC
2 years, 0 months	94[b,c]	87[b,c]	—	—
2 years, 6 months	87[b,c]	69[c]	—	79[b,c]
3 years, 0 months	73[b,c]	57[c]	62[c]	70[b,c]
3 years, 6 months	66[c]	47	57[c]	60[c]
4 years, 0 months	55[c]	40	48	60[c]
4 years, 6 months	50	31	45	54
5 years, 0 months	44	27	43	58[c]
5 years, 6 months	41	24	42	55[c]

Note. M = 100, SD = 16 for SB-LM and SB4; M = 100, SD = 15 for the WPPSI and K-ABC.

[a]SB4 Composites are based on the assumption that a valid score (i.e., raw score > 1) is obtained on each subtest appropriate for administration at a given age level and ability level.

[b]Principal indicator is insensitive to performances more than two standard deviations below the test mean.

[c]Principal indicator is insensitive to performances more than three standard deviations below the test mean.

TABLE 10.3. Maximum Overall Ability Scores for Select Age Groups on SB4, SB-LM, the Wechsler Scales, and the K-ABC

Age in years and months	SB4[a]	SB-LM	Wechsler scale[b]	K-ABC
2 years, 0 months	164	162	—	—
4 years, 0 months	164	160	160	160
6 years, 0 months	164	159	160	160
8 years, 0 months	164	164	160	160
10 years, 0 months	164	160	160	160
12 years, 0 months	164	164	160	155
14years, 0 months	158	154	160	—
16 years, 0 months	152	138	155	—
18 years, 0 months	149	136	155	—
20 years, 0 months	149	—	155	—

Note. $M = 100$, $SD = 16$ for SB4 and SB-LM; $M = 100$, $SD = 15$ for all Wechsler scales and the K-ABC.
[a]For any given age level, SB4 Composites are based on the maximum number of subtests specified in Appendix F of the *Guide for Administering and Scoring* (Thorndike et al., 1986a).
[b]The WPPSI-R Full Scale IQ (FSIQ) is the principal Wechsler indicator at age 4 years, 0 months; the WISC-III FSIQ is used at ages 6 years, 0 months through 16 years, 0 months; and the WAIS-III FSIQ is used at ages 18 years, 0 months and 20 years, 0 months.

clusion of standard scores three or more standard deviations above the test mean, SB4 discriminates talent as adequately as SB-LM did at all age levels, and it possesses slightly higher ceilings at ages 16 and above (columns 2 and 3). The Composite also compares favorably to optimal performance on the Wechsler scales and the K-ABC (columns 4 and 5, although the latest revisions of the Wechsler scales have eliminated most of SB4's previous advantage in this area). These comparisons suggest that the SB4 would be a good choice for evaluations assessing potentially gifted youths, although it provides significantly higher scores than the more recent WISC-III (Simpson et al., 2002).

STANDARDIZATION

The goal in developing the standardization sample for the SB4 was to approximate the demographics of the United States based on the 1980 census (Thorndike et al., 1986c). There have been important demographic changes in the two decades since then. Most notable has been the increase in ethnic minority populations, particularly Spanish-speaking groups (Hernandez, 1997). Two interrelated issues must be considered in regard to the representativeness of SB4 norms. The first is the loss of randomness that resulted from the need to obtain examinees' cooperation. The second is the weighting of test scores to compensate for discrepancies between the designated sampling plan for socioeconomic status (SES) and SES levels in the obtained sample.

Nonrandomness and General Referents

One popular view holds that the strength of an IQ test depends upon the degree to which its sample represents the general population. "Stratified random sampling" would be a relatively efficient method for obtaining such a representation. Many practitioners, as well as notable measurement specialists (e.g., Hopkins, 1988), assume that individually administered IQ tests are normed on stratified random samples. This, however, is never the case. Test developers must request examinees' cooperation. The net effect is a loss of randomness, because people who volunteer are rarely like those who do not (Jaeger, 1984).

The common alternative to stratified random sampling is to select examinees purposively through "quota sampling." The shortcoming of quota sampling is that its selections are likely to be biased, unless of course cooperation rates are high and uniform across strata (Hansen, Hurwitz, & Madow, 1953; Kish, 1965; Thorndike, 1982). SB4 was normed on 5,013 individuals arranged into 17 age groups (2 years, 0 months through 23 years, 11 months). Quo-

ta sampling was employed to approximate the U.S. population in terms of geographic region, community size, race, gender, and SES. Unfortunately, lower-SES examinees were underrepresented in the sample (10.6% vs. 29.2% of the U.S. population), and higher-SES examinees were overrepresented (43.1% vs. 19.0%, respectively).

It would be simplistic to discredit SB4 for sampling problems. The quota sampling in SB4, as well as the differential rates of cooperation, are common to *all* individually administered IQ tests, including the K-ABC, WISC-III, and the Woodcock–Johnson Psycho-Educational Battery—Revised (WJ-R) Tests of Cognitive Ability (Woodcock & Johnson, 1990a). The standardization samples of even the best available instruments are imperfect approximations of the general U.S. population at the time any given instrument was developed.

Nonrandomness and Other Referents

An alternative perspective is that it is not necessary for IQ tests to reference the general population. There are other legitimate referents to which test scores can be compared. Instruments such as SB4 are most often administered to two groups—namely, examinees who truly volunteer to be tested and examinees with suspected disabilities. Consequently, it is essential that IQ tests accurately reflect the capabilities of these two groups. Examinees who willingly consent to testing (self-referred individuals, those who may be gifted, certain segments of the population receiving special education) do not necessarily differ from the "volunteer" subjects in standardization samples. At least in this regard, IQ test norms should be appropriate for volunteers. The second group is more problematic. Individuals with disabilities are administered IQ tests for special purposes (e.g., assignment to special education categories, mandatory reevaluations). As such, most of these individuals cannot be truly regarded as volunteers. Clearly linked to this phenomenon is the need to consider persons with disabilities *systematically*—if not directly in test norms, then through special studies.

One proposal for test development is to sample individuals with disabilities in proportion to their presence in the general population. Such an approach assumes that prevalence rates are known for the various exceptionality subtypes. This assumption is problematic for such conditions as learning disabilities, for which there is no uniformly accepted rate of occurrence in the general population and for which diagnostic rates continue to escalate (see Ysseldyke & Stevens, 1986). The dilemma in such instances becomes this: "What is the appropriate percentage of individuals with learning disabilities to include in stadardization samples?"

Unsettling problems arise even when prevalences are known. A prevalence of 3% is the standard endorsed for mental retardation (Grossman, 1983). Yet it would be improper to systematically target individuals identified as having mental retardation to form 3% of a test's sample. Probability theory dictates that a percentage of the volunteers in the sample who are *not* thus identified will also have mental retardation. When the two groups are merged, individuals with mental retardation will be overrepresented. Counterintuitively, this overrepresentation increases the likelihood that test norms will be diagnostically insensitive to persons with mental retardation. The overrepresentation of low-scoring examinees (i.e., those with retardation) will affect the conversion of raw scores to normalized standard scores. As a result, a *lower* raw score will be needed to obtain an IQ in the range for mental retardation (i.e., an IQ < 70) than if such examinees had not been oversampled. The diagnostic result is that test norms will fail to qualify higher-functioning individuals with mental retardation for needed services.

One final issue is that either approach—whether including specific conditions in the standardization sample, or developing separate specialized samples—assumes that developers will include all potentially relevant diagnostic categories. Unfortunately, at present we have incomplete knowledge of the different exceptionalities that might influence performance on a cognitive ability battery. There is evidence that some psychiatric diagnoses, such as attention-deficit/hyperactivity disorder (e.g., Saklofske, Schwean, Yackulic, & Quinn, 1994; Schwean, Saklofske, Yackulic, & Quinn, 1993) or autism (Carpentieri & Morgan, 1994; Harris, Han-

dleman, & Burton, 1990), may be associated with mean differences in performance on at least some aspects of ability. Currently it is unclear whether these group differences reflect changes in cognitive processing, or whether the effect is mediated by changes in motivation or test session behavior (Glutting, Youngstrom, Oakland, & Watkins, 1996).

Thus there are at least four links in the chain connecting knowledge to the design of an appropriate standardization sample: (1) awareness of all the conditions and exceptionalities that may influence performance on an ability test; (2) accurate data about the prevalence rate of these conditions in the population; (3) efficient and affordable ways of identifying potential participants meeting criteria for the conditions, either by doing so among the "volunteers" or by generating a special reference sample; and (4) a clear theoretical rationale about the appropriateness of developing a separate set of norms for a particular group (e.g., is it meaningful to know how the working memory performance of a youth with depression compares to other such youths, or only to other youths the same age, regardless of exceptionality?). Given these hurdles, the most practical solution for test developers will probably continue to be approximation of stratified sampling, with the hope that participation biases do not lead to serious underrepresentation of important conditions. A statistical alternative might be to explicitly model the selection process for participants, and then use estimates based on this model to correct observed values for "nonsampling bias" (see Wainer, 1999, for discussion and examples). Either way, it is important for test consumers and users to remain aware of these assumptions about the representativeness of the standardization sample.

Enhancing Diagnostic Utility

For the reasons discussed above, proportional sampling of individuals with disabilities is likely to create as many problems as it solves. A more practical response is to systematically oversample these individuals, but not necessarily to include them in test norms. Instead, special studies should be conducted to determine how the test behaves in these populations. Confirmatory factor analysis, for example, could identify whether test dimensions are similar for persons with and without disabilities (e.g., Keith & Witta, 1997). Comparisons based on item response theory (IRT; e.g., Hambleton, Swaminathan, & Rogers, 1991) could verify whether item difficulties are identical among exceptional and nonexceptional groups. IRT would also uncover whether item calibrations are sufficient for the maximum differentiation of low-scoring and high-scoring exceptionalities (Embretson, 1999). Multiple-regression slope comparisons (and not bivariate correlations) should supply information relevant to whether test scores predict equally for persons with and without disabilities (Jensen, 1980). Finally, univariate and multivariate contrasts could shed light on whether all possible test scores (e.g., overall IQs, factor scores, subtest scores) differ between the general sample and the various exceptionality subtypes (persons with mental retardation, learning disabilities, etc.).

Compared to these ideals, SB4 leaves room for improvement. This finding is disappointing, since sufficient data were gathered during SB4's development to complete many of the analyses identified above. SB4 is to be commended for verifying that its Composite and area scores (but not necessarily subtest scores) differ between exceptional and nonexceptional samples. Nevertheless, no attempt was made to determine whether SB4's items are unbiased for those with disabilities, or that its test dimensions are similar for individuals functioning normally and exceptionally. Likewise, although criterion-related validity is reported for those with disabilities, quantitative comparisons were not conducted for the relative accuracy of predictions between those with and without disabilities.

In fairness, no IQ test has met all of these standards at the time of its publication. However, the issue is not whether SB4 should be excused because it is no more deficient than other ability tests. Rather, the issue is why IQ tests are marketed without adequate evidence that they reflect the aptitudes of individuals with disabilities. We, as professionals responsible for the welfare of clients, must demand this information at the time of a test's publication. Otherwise, we

must accept the fact that we are willing to apply tests whose diagnostic capabilities are unknown.

Elimination versus Weighting

There is often "slippage" between a test's sampling plan and the testing as executed (Thorndike, 1982). Two methods can bring a sample back into alignment with its sampling plan. The first option is to eliminate examinees randomly from oversampled strata. The second option is to *weight* scores from each stratum to their correct percentage of the population. Whereas both methods have their benefits, neither can fully compensate for a loss of randomness in the sampling process (Glutting & Kaplan, 1990).

Until SB4, norms for individually administered IQ tests were typically aligned by eliminating examinees from oversampled strata. The benefit of elimination is that there is no redundancy in subject-generated variance (i.e., an examinee is not counted as more, or less, than one case). Moreover, the practice is tidy. "Final" samples often align well with the population, in part, because test manuals provide little discussion of discarded cases. Therefore, had SB4 used elimination, it would have been easy to marvel at how well the sample approximated the general population on race, gender, SES, and so on. Instead, SB4 retained all 5,013 participants in the standardization sample, even though higher-SES families were more likely to provide informed consent and to participate than were lower-SES families. In an effort to correct for these sampling biases, the SES variables of occupation and education were weighted so that examinees' scores would conform to their correct percentages in the U.S. population. That is, "each child from an advantaged background was counted as only a fraction of a case (as little as 0.28), while each child from a less advantaged background was counted as more than one case" (Thorndike et al., 1986c, p. 24).

One advantage of weighting is that it accounts for all scores in the sample. Relatedly, it produces estimates of higher reliability than does elimination. A potential flaw is that weighted estimates are not based entirely on actual cases. Examinees in underrepresented strata are counted *more than once* by multiplying the original sample variance upward to the desired population estimate. The process is dependent upon the assumption that examinees in the sample are representative of the *entire* population—including those individuals who, for whatever reason, were not sampled.

There is no guarantee that the scores of examinees already in the sample are similar to the scores of potential examinees who were not tested. However, this assumption becomes more plausible when the obtained sample has large numbers of examinees in each stratum who are representative of that particular population segment. SB4's standardization sample is quite large (n = 5,013), and its strata are probably of sufficient subject size for weighting. Moreover, weighting is an accepted procedure for standardizing group tests. Consequently, from this perspective, the weighting of test scores in SB4 appears as reasonable as the weighting used to norm group tests.

RELIABILITY

By and large, SB4's reliabilities are quite good. Internal consistency for the Composite is excellent, with Kuder–Richardson 20 coefficients ranging from .95 to .99 across age levels. Reliabilities for area scores are also substantial. Internal consistency for two-, three-, and four-subtest groupings vary from .86 to .97 for Verbal Reasoning (median r = .95). Coefficients for Abstract/Visual Reasoning range from .85 to .97 and show a median of .95. Similarly, estimates for Quantitative Reasoning vary from .80 to .97 (median r = .94), and internal consistency for Short-Term Memory ranges from .86 to .95 (median r = .86). It is worth noting that only the Composite achieves reliability coefficients consistently greater than Kelley's (1927) recommended threshold of .94 for making decisions about individuals. Most of the area scores attain the less conservative threshold (reliabilities $\geq$ .85) proposed by Weiner and Stewart (1984) for individual classification.

Subtest internal consistencies are lower, as would be expected from their shorter test lengths. Nonetheless, with the exception of one subtest, median coefficients are reason-

ably high (range = .83 to .94 across age groups). The exceptional subtest (Memory for Objects) is located in the Short-Term Memory area, and it produces coefficients of marginal reliability (median r = .73). The subtest with the second lowest reliability is also in the Short-Term Memory area (Memory for Digits; median r = .83). As a result, psychologists should be alert that subtest scores from the Short-Term Memory area are likely to be less precise than subtest scores from other areas in SB4.

Standard errors of measurement (SEMs), and "confidence bands" derived from SEMs, are the reliability issues most likely to affect everyday practice. Confidence bands produce information relevant to the fallibility of test scores, and consequently help to clarify the relative verity and utility of test scores in decision making about individuals (Glutting, McDermott, & Stanley, 1987).

Memory for Objects provides the least precise scores in SB4 (i.e., the largest confidence bands). Its SEM shows a median of 4.15 points across age groups. The subtest with the second largest SEM is Memory for Digits (median = 3.25). However, the SEMs of these two subtests (and for all other more reliable subtests) are within reasonable limits. Also, as might be expected, greater precision in scores is found when interpretations are based on the four area scores. Median SEMs for Verbal Reasoning, Abstract/Visual Reasoning, Quantitative Reasoning, and Short-Term Memory are as follows: 3.9, 3.6, 3.8, and 4.8, respectively. Finally, the most precise score in SB4 is the Composite (median SEM = 2.3; all SEMs as reported in Sattler, 1992).

The *Technical Manual* (Thorndike et al., 1986c) calculates score stability for samples of preschoolers (5-year-olds) and children attending elementary school (8-year-olds). Preschoolers' test–retest coefficients are reasonable for the Composite (r = .91) and for area scores (range = .71 to .78). Less stability is evident for individual subtests, and in particular for Bead Memory (r = .56). The pattern of test–retest coefficients of elementary school children is similar to that found for preschoolers. Appreciable stability is present for the Composite (r = .90) and for the areas of Verbal Reasoning, Abstract/Visual Reasoning, and Short-Term Memory (r's = .87, .67, and .81, respectively). However, somewhat lower stability is found for the Quantitative Reasoning area (r = .51).

Preschoolers' Composites will, on average, increase approximately 8.2 points from test to retest administrations. Similarly, Composites are likely to increase by 6.4 points for elementary school children who are tested twice across short intervals. SB4 offers no stability data for examinees of junior high or high school age, or for young adults, making it difficult to approximate the score increases that might be expected of these age groups.

VALIDITY

An impressive amount of validity information has been gathered in support of SB4. In particular, investigations have addressed developmental changes of raw scores by age; quantitative analyses of item fairness across gender and ethnic groups; correlations with other IQ tests, using samples of both normal and exceptional examinees; correlations with achievement tests; score differences between the standardization sample and special groups (individuals who are gifted, have learning disabilities, or have mental retardation); and the factor structure of SB4's test dimensions.

Concurrent Validity

Table 10.4 presents concurrent correlations between SB4 and other IQ tests administered to normal samples. This compilation was obtained from studies reported in the *Technical Manual* (Thorndike et al., 1986c) and in a review by Laurent, Swerdlik, and Ryburn (1992), as well as from studies conducted by independent investigators. Laurent and colleagues present validity data for exceptional samples, too. Results show substantial associations between SB4's Composite and overall scores on SB-LM, all Wechsler scales, the K-ABC, the WJ-R (Woodcock & Johnson, 1990a), and the Differential Ability Scales (DAS; Elliott, 1990). Correlations ranged from .53 to .91 (average r = .78 using Fisher's z' transformation). The consistency and magnitude of these relationships speak well for the Composite's construct validity.

Spearman's (1923) principle of the "indif-

TABLE 10.4. Score Characteristics and Correlations of SB4 with Other IQ Tests Administered to Nonexceptional Samples

Study	*n*	Mean age (years)	Mean SB4 Composite	Other IQ test	Other test mean IQ	IQ difference	Correlation
Elliott (1990)	55	9.9	109.8	DAS	106.3	3.5	.88
Thorndike, Hagen, & Sattler (1986c, Study 5)	175	7.0	112.7	K-ABC	112.3	0.4	.89
Hendershott, Searight, Hatfield, & Rogers (1990)	36	4	110.5	K-ABC	118.2	−7.7	.65
Krohn & Lamp (1989)[a,b]	89	4.9	93.4	K-ABC	96.0	−2.6	.86
Krohn & Lamp (1989)[a,b]	65	4	93.8	K-ABC	95.8	−2.0	—
Krohn & Lamp (1989)[a,b]	65	6	93.3	K-ABC	99.7	−6.4	—
Krohn & Lamp (1989)[a,b]	65	9	96.5	K-ABC	97.9	−1.4	—
Kaufman & Kaufman (1983)	121	School-age	116.5	K-ABC	114.5	+2.0	.61
Smith & Bauer (1989)	30	4.9	—	K-ABC	—	—	.57
Clark, Wortman, Warnock, & Swerdlik (1987)	47	—	—	SB-LM	—	—	.53
Hartwig, Sapp, & Clayton (1987)	30	11.3	113.1	SB-LM	114.4	−1.3	.72
Thorndike, Hagen, & Sattler (1986c, Study 1)	139	6.9	105.8	SB-LM	108.1	−2.3	.81
Krohn & Lamp (1989)[a,b]	89	4.9	93.4	SB-LM	—	—	.69
Lukens (1988)	31	16.75	44.8	SB-LM	46.7	−1.9	.86
Psychological Corporation (1997)	26	28.6	114.8	WAIS-III	113.3	+1.5	.88
Carvajal, Gerber, Hughes, & Weaver (1987)	32	18.0	100.9	WAIS-R	103.5	−2.6	.91
Thorndike, Hagen, & Sattler (1986c, Study 4)	47	19.4	98.7	WAIS-R	102.2	−3.5	.91
Lavin (1996)	40	10.6	108.0	WISC-III	107.0	+1.0	.82
Rust & Lindstrom (1996)	57	6–17	109.9	WISC-III	111.3	−1.4	.81
Rothlisberg (1987)	32	7.8	105.5	WISC-R	112.5	−7.0	.77
Thorndike, Hagen, & Sattler (1986c, Study 2)	205	9.4	102.4	WISC-R	105.2	−2.8	.83
Carvajal & Weyand (1986)	23	9.5	113.3	WISC-R	115.0	−1.7	.78
Greene, Sapp, & Chissom (1990)[c]	51	Grades 1–8	80.5	WISC-R	78.1	+2.4	.87
Wechsler (1991)	205	6–16	—	WISC-R	—	—	.83
Woodcock & Johnson (1990b, Study 1)	64	2.9	—	WJ-R	—	—	.69
Woodcock & Johnson (1990b, Study 2)	70	9.5	—	WJ-R	—	—	.69
Woodcock & Johnson (1990b, Study 3)	51	17.5	—	WJ-R	—	—	.65
Thorndike, Hagen, & Sattler (1986c, Study 3)	75	5.5	105.3	WPPSI	110.3	−5.0	.80
Carvajal, Hardy, Smith, & Weaver (1988)	20	5.5	114.4	WPPSI	115.6	−1.2	.59

(*continues*)

TABLE 10.4. *Continued*

Study	*n*	Mean age (years)	Mean SB4 Composite	Other IQ test	Other test mean IQ	IQ difference	Correlation
Carvajal, Parks, Bays, & Logan (1991)	51	5.7	103.0	WPPSI-R	109.5	−6.5	.61
McCrowell & Nagle (1994)	30	5.0	95.9	WPPSI-R	94.1	+1.8	.77
Wechsler (1989)	105	5.6	107.2	WPPSI-R	105.3	+1.9	.74

Note. DAS, Differential Ability Scales; K-ABC, Kaufman Assessment Battery for Children; SB-LM, Stanford–Binet, Form L-M; WAIS-R, Wechsler Adult Intelligence Scale—Revised; WAIS-III, Wechsler Adult Intelligence Scale—Third Edition; WISC-R, Wechsler Intelligence Scale for Children—Revised; WISC-III, Wechsler Intelligence Scale for Children—Third Edition; WPPSI, Wechsler Preschool and Primary Scale of Intelligence; WPPSI-R, Wechsler Preschool and Primary Scale of Intelligence—Revised; WJ-R, Woodcock–Johnson Psychoeducational Battery—Revised.
[a]Same sample appears multiple times in table, because participants completed multiple ability tests.
[b]Head Start sample, followed longitudinally.
[c]African American sample.

ference of the indicator" suggests that the specific item content in intelligence tests is unimportant to the evaluation of general ability (or *g*). The truly important phenomenon for *g* is that IQ tests measure inductive and deductive reasoning. Thus correlations between one IQ test and IQ tests with dissimilar content can help evaluate the extent to which the first test measures *g*. Based on the correlations in Table 10.4, at least 60.8% of the Composite's variance is accounted for by *g*. These data suggest that the Composite provides a reasonably trustworthy estimate of general intelligence.

Of applied interest are score differences that can be expected between SB4 and other IQ tests. Column 7 in Table 10.4 (labeled "IQ difference") shows that the Composite *averages* 2.5 points lower than the IQs from other intelligence tests published prior to SB4. Interestingly, scores on the SB4 also average 0.5 points *higher* than scores obtained on tests published after SB4 (i.e., the WPPSI-R, DAS, WJ-R, WISC-III, and WAIS-III). Individual comparisons are less precise because of the smaller number of studies between SB4 and any one test. With this caveat in mind, psychologists might expect SB4 to produce IQs that are about 2 points lower than those from SB-LM; 5 points lower than those from the WISC-R; 3 points lower than those from the Wechsler Adult Intelligence Scale—Revised (WAIS-R); 3 points lower than those from the WPPSI; and 2.5 lower than those from the K-ABC. Given the difference in times when these respective standardization samples were collected, it is likely that the "Flynn effect" accounts for much of this variance in scores (Flynn, 1984, 1999). By virtue of the Flynn effect, which refers to apparent secular gains in average performance on ability tests, it is likely that SB4 scores would be about 3 points higher than scores derived from tests normed a decade later, such as the WAIS-III, the forthcoming revision of the K-ABC, and the new WJ-III.

Factor Structure

The most controversial aspect of SB4 concerns the interpretability of its area scores. That is, do the capabilities evaluated by SB4 actually conform to the four-factor model of intelligence that has been advanced for the test? This question of construct validity is open to empirical verification, and it is one that usually can be answered through factor analysis.

It is disconcerting that the authors of SB4 themselves disagree about the number of interpretable factors. Thorndike, for example, in light of his own factor-analytic results, offers no explanation for why the four-factor model should be applied to examinees younger than age 12. He "confirms" only two factors between ages 2 and 6 (Verbal, Abstract/Visual). His analyses then support a three-factor model between ages 7 and 11 (Verbal, Abstract/Visual, Memory). Most importantly, the proposed four-factor model does not emerge until ages 12 through 23.

Sattler (1992), on the other hand, eschews the SB4 model. He proposes a two-

factor solution between ages 2 and 6 (Verbal Comprehension, Nonverbal Reasoning/Visualization) and a three-factor solution at ages 7 through 23 (Verbal Comprehension, Nonverbal Reasoning/Visualization, Memory). Conspicuously absent in Sattler's findings is the dimension of Quantitative Reasoning, and at no age level does he recommend the interpretation of all four area scores.

Perhaps because of this open disagreement, investigators have extensively reanalyzed the SB4 normative data as well as conducting independent replications. We found a dozen different published factor analyses of SB4. Four provided evidence consistent with the published four-factor structure (Boyle, 1989, 1990 [especially if one is willing to exclude certain subtests]; Keith, Cool, Novak, & White, 1988; Ownby & Carmin, 1988). Five studies challenge the four-factor structure, suggesting anywhere from one general ability factor (Reynolds, Kamphaus, & Rosenthal, 1988) to two or three factors, depending on age (Gridley & McIntosh, 1991; Kline, 1989; Molfese, Yaple, Helwig, & Harris, 1992; Sattler, 1992). The remaining studies are equivocal about the competing models (McCallum, Karnes, & Crowell, 1988; Thorndike et al., 1986c; Thorndike, 1990). There is a tendency to detect more factors in older age groups, with a two-factor structure describing preschool data, and three factors describing data for youths above 7 years of age. These differences in factor structure, if true, could reflect either developmental change or alterations in the subtest battery administered at each age. Older youths completed more subtests on average, increasing the likelihood of statistically recovering additional factors (if such additional dimensions of ability were measured by the subtests).[3]

Interestingly, we are not aware of any published study that has used either Horn's parallel analysis (Horn, 1965) or the method of minimum average partials (Velicer, 1976) as decision rules to determine the appropriate number of factors to retain for the SB4. Methodological evidence strongly suggests that these are the two techniques most likely to recover the accurate number of factors, and they tend to retain fewer factors than more commonly used procedures, such as the maximum-likelihood chi-square test or the Kaiser criterion (Zwick & Velicer, 1986). The common element in all this is that no single study has definitively substantiated the existence of *four* area factors. It therefore stands to reason that psychologists should refrain from interpreting area scores until more evidence is offered on their behalf.

It should be kept in mind that current inabilities to support a four-factor model may not necessarily represent a failure of SB4 per se. Rather, the difficulty may lie in the sensitivity of factor analysis to data-related issues in SB4. This is particularly true when confirmatory factor analysis is applied. The relationship between confirmatory factor analysis and SB4 was explored in detail by Glutting and Kaplan (1990).

SCORE INTERPRETATION

The SB4 can potentially support clinical interpretation at a variety of levels of analysis. The battery yields a single, global estimate of general cognitive ability, the Composite score, which represents the most general level of analysis available on the SB4. Beneath the Composite, the SB4 also theoretically could yield scores for Fluid and Crystallized cognitive ability, which are referred to as "Level II scores" in the SB4 manuals. SB4 also includes the Short-Term Memory factor score in Level II. Level III scores on the Binet include factor-based scores measuring the specific cognitive abilities of Verbal Reasoning, Quantitative Reasoning, and Abstract/Visual Reasoning. SB4, unlike SB-LM, also provides standardized age scores for specific subtests, enabling potential interpretation of subtest profiles. This exemplifies the most fine-grained level of clinical interpretation that would be considered in most cases (cf. Sattler, 1992, for discussion of attention to responses to specific items).

This structure for the SB4 is similar to the hierarchical structures adopted by most contemporary measures of cognitive ability, and this format lends itself readily to the "top-down" models of interpretation advocated by many assessment authorities (e.g., Aiken, 2000; Kamphaus, 1993; Kaufman, 1994; Sattler, 1992). It is important to consider the evidence supporting these different levels of interpretation; assessment practice

is better driven by scientific evidence than by convention and appeals to authority.

The Level I score, the Composite, possesses good evidence of validity. The preponderance of research involving the SB4 and its predecessors has concentrated on the Composite score, so there is considerable accumulated evidence about the Composite score's convergent, criterion, and predictive validity for such constructs as academic achievement. The Composite score also has gained fairly consistent support from factor analyses of the SB4 subtests (which typically have indicated either several correlated factors or one general ability factor). Although some have questioned the treatment validity of even these most global scores from cognitive ability tests (Macmann & Barnett, 1997; McCallum et al., 1988), a good case can be made for using and interpreting these global scores (Neisser et al., 1996), particularly in terms of psychoeducational and vocational assessment.

Level II scores are less well supported. The SB4 as published provides an observed score for the Short-Term Memory area, but the Abstract/Visual Reasoning area score is the only potential indicator of Fluid-Analytic Abilities in the battery. This limits the construct validity of estimates of fluid ability derived from the SB4, inasmuch as fluid ability may involve processes beyond abstract/visual reasoning (Carroll, 1993; Horn & Noll, 1997). Furthermore, the SB4 manuals and interpretive aids do not formally present a way of calculating a summary score for Crystallized Abilities, although it is possible to estimate such a score by combining the Verbal and Quantitative Reasoning area scores. The proposed Level II structure of the SB4 has not consistently been confirmed by secondary analyses of the standardization data or independent samples. Perhaps most crucially, there is a dearth of research addressing the criterion validity of Level II scores from the SB4 (cf. Caruso, 2001). The paucity of research is probably largely related to the lack of emphasis on Level II interpretation in the SB4 materials, and it is possible that future research will demonstrate value in interpreting these more discrete ability estimates (e.g., Moffitt & Silva, 1987). At present, however, there is minimal literature to guide clinical hypothesis generation or interpretation of Level II scores, and there is little guidance offered to the practitioner about how to calculate the Level II scores beyond Short-Term Memory. This level of analysis has not received much attention in practice, and it probably should not be emphasized until more evidence is available demonstrating clear incremental validity above and beyond the information derived from the Composite score.

Level III scores are also problematic, because of disagreement about factor structure as well as a lack of information about incremental validity. The purported structure of Verbal Reasoning, Quantitative Reasoning, and Abstract/Visual Reasoning as three distinct factors has not consistently emerged across the ages covered by SB4, or in analyses of independent samples (and not always in secondary analyses of the standardization data). Currently there is insufficient evidence to permit us to conclude whether the subtests on the SB4 adequately assess these three different dimensions of ability. More importantly from a practical perspective, at present there is no information about incremental validity for these area scores after the Composite score is interpreted. Although researchers have begun to explore the possibility that more discrete ability scores might provide additional clinical data about achievement or behavior problems not subsumed in a more global ability score (cf. Glutting, Youngstrom, Ward, Ward, & Hale, 1997; Youngstrom, Kogos, & Glutting, 1999), this work still needs to begin with the SB4. Area score interpretation imposes burdens on the practitioner and consumer in terms of longer tests, greater complexity of results, and potentially greater likelihood of diagnostic errors (Silverstein, 1993). In light of these costs, it would seem premature to emphasize Level III area scores in SB4 interpretations. The lack of consensus about the construct validity for these scores, based on factor analyses, further calls for caution.

Psychologists may be tempted to make "area" interpretations (e.g., Naglieri, 1988a), even though there is little justification for this practice. Indeed, Hopkins (1988) appears to believe that SB4's four area scores should be interpreted, and that practitioners need not "become emotionally involved in the 'great debate' regarding the theoretical structure of intelligence a deter-

mined by the factor analytic method" (p. 41). Hopkins's position is incorrect, because it implies that clinical necessity should supersede what can be supported empirically. However, the need to generate hypotheses about an examinee is *never* sufficient grounds for the interpretation of a test score. This is especially true in the case of SB4's four area scores, since claims for their construct validity have yet to be substantiated, in spite of a considerable amount of investigation. Even if this were accomplished, it would also be necessary to document criterion validity and incremental validity (above more parsimonious *g*-based models) before clinical interpretation of area scores could be justified.

The addition of standard age scores for subtests to the SB4 created the possibility of subtest profile interpretation, which has become a prevalent practice in the use of other major ability tests (e.g., Kaufman, 1994; Naglieri, 1988b; Sattler, 1992). Many clinicians and researchers welcomed this addition as an opportunity to improve the perceived clinical value of the SB4, hoping that more detailed attention to patterns of performance on subtests would lead to improved psychoeducational prescription (Lavin, 1995) or to identification of profiles characterizing the performance of specific diagnostic groups (e.g., Carpentieri & Morgan, 1994; Harris et al., 1990). Procedures and recommendations are available to promote this sort of analysis with the SB4 (Naglieri, 1988b; Rosenthal & Kamphaus, 1988; Spruill, 1988). Sattler (1992) also provides a detailed table (Table C-52) listing the abilities thought to be reflected in each subtest, background factors thought to affect subtest performance, possible implications of high and low scores on each subtest, and instructional implications of unusual performance on each subtest. Sattler's table is thorough. For example, Sattler lists from 3 to 18 distinct abilities for each of the 15 subtests ($M = 8.5$, $SD = 3.8$), and an average of five implications for every high or low score per subtest. This presentation clearly encourages the clinical interpretation of individual strengths and weaknesses at the subtest level. Although Sattler provides some cautionary statements about not interpreting a subtest score in isolation, such tables seem prone to abuse. The situation confronting the clinician is complex: Who can generate a hypothesis with any confidence when faced with an average of eight or nine different abilities and another three background factors that could contribute to performance on a specific subtest?

In addition, subtest analysis faces substantial psychometric challenges (Macmann & Barnett, 1997; McDermott, Fantuzzo, & Glutting, 1990; McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992) that make it unlikely to deliver on the promise of improved assessment or treatment planning. In fact, the studies available to date for the SB4 clearly indicate that there is no significant improvement in assessment when subtest interpretation is added to the analytic strategy (Kline, Snyder, Guilmette, & Castellanos, 1992, 1993). This is consistent with growing evidence from investigations with other tests, indicating that subtest analysis is problematic at best when applied to routine assessment goals such as predicting academic achievement or diagnosis (e.g., Watkins, 1996; Watkins, Kush, & Glutting, 1997). In short, it appears that the SB-LM was not missing much by failing to include subtest scores, and that practitioners would do well to avoid relying much on SB4 subtests as a distinct level of analysis in conducting evaluations.

SHORT FORMS

Cognitive assessment is a time-consuming enterprise (Meyer et al., 1998). This expense, combined with the lack of validity information supporting the clinical use of scores beyond the Composite as described above, strongly suggests the potential value of short forms of the SB4 that provide reliable estimates of general ability without entailing the costs of a complete administration. SB4 offers four short forms that result in a substantial savings of testing time: the six-subtest General Purpose Assessment Battery (GPAB; Vocabulary, Bead Memory, Quantitative, Memory for Sentences, Comprehension, and Pattern Analysis); the four-subtest Quick Screening Battery (Vocabulary, Bead Memory, Quantitative, and Pattern Analysis); the four- to six-subtest Battery for the Assessment of Students for Gifted Programs; and the six-subtest Battery

for Students Having Problems Learning in School. Short forms with four or fewer subtests are intended for screening purposes, but batteries composed of at least six subtests can be used for placement decisions (Thorndike et al., 1986c, p. 50). This latter possibility makes it essential that test scores from six-subtest abbreviated batteries be psychometrically equivalent to those from the full test.

According to the data presented in the *Technical Manual* (Thorndike et al., 1986c), split-half reliabilities for two-, four-, and six-subtest short forms are fairly constant and appreciable for examinees of different ages. Correlations between Composites from short forms and the complete battery are also acceptable. However, the *Technical Manual* fails to present information about differences between estimated Composites and area scores from abbreviated batteries and actual scores on the full test. Since publication of the SB4, more than a dozen independent studies have investigated the psychometric properties of various abridged forms, using samples ranging from low to high ability and from preschool to college. The majority of these investigations concluded that the six-subtest GPAB was the most acceptable substitute for a complete administration (Atkinson, 1991; Carvajal & Gerber, 1987; DeLamatre & Hollinger, 1990; Kyle & Robertson, 1994; McCallum & Karnes, 1990; Prewett, 1992; Volker, Guarnaccia, & Scardapane, 1999), possessing both good correspondence with the full battery and good external validity with other measures of ability (Carvajal, Hayes, Lackey, & Rathke, 1993; Carvajal, McVey, Sellers, & Weyand, 1987). On the other hand, two investigations concluded that the four-subtest battery performs essentially as well as the six-subtest version, and argued that the four-subtest version is preferable for screening purposes in light of its brevity (Prewett, 1992; Volker et al., 1999). Finally, Nagle and Bell (1993) found that all of the short forms produced what they considered to be unacceptable levels of disagreement for individual classification purposes. Instead, these authors recommend the use of item reduction short forms rather than subtest reduction versions (Nagle & Bell, 1995). On the whole, these studies alleviate earlier concerns that the short forms might show substantially lower external validity, in spite of correlating well with the full-battery composite (Levy, 1968; McCormick, 1956). It is less clear that short forms provide an adequate substitute for the full battery when individual classification decisions are required; in addition, the two-subtest battery clearly is suitable only for group research and not individual assessment.

In spite of the burgeoning literature examining SB4 short forms, important questions remain unanswered. One problem is that practitioners often may develop idiosyncratic short forms that have not been empirically validated. Norms tables in SB4 make it possible to calculate Composites from practically *any* combination of subtests. Thus practitioners can develop their own short forms by "picking and choosing" among favorite subtests. No matter how the particular combination is chosen, problems are likely to arise for short forms if the *administration sequence* of the subtests is disturbed.[4] Assume, for example, that a psychologist elects to administer a short form consisting of subtests 1, 3, 4, 5, 6, and 13. The psychologist in such an instance is operating under the belief that norms for subtest 13 (Paper Folding and cutting) will remain constant, regardless of the fact this subtest now occupies position 6 in the new battery. Thus the validity of the procedure is critically dependent on the assumption that norms and examinees' performances are independent of a subtest's location in the battery.

Such assumptions of independence are certainly open to question. Decreases in testing time may lessen an examinee's frustration and improve test scores on the shorter battery. Differences in the fatigue of the psychologist or examinee, or the fact that the full test offers more opportunities to gain experience in understanding test directions and familiarity with test materials, could also affect performance. Learning or "carryover" effects from one subtest to the next are particularly likely for measures that require examinees to manipulate objects (i.e., nonverbal/performance subtests). Finally, even if these assumptions were satisfied, the psychologist must consider whether the external validity of the shorter battery is the same as that of the full test

(see Smith, McCarthy, & Anderson, 2000, for further recommendations about the development and evaluation of short forms). This limitation also applies to the majority of extant research with SB4 short forms: Researchers typically administer the full battery and then extract different short forms from that battery. Thus practice, fatigue, and motivation effects are based on a full administration, which would not be the case when a short form was administered clinically.

Without more information about the effects of subtest sequencing and battery length, as well as short-form external validity, it could be argued that psychologists should administer SB4 in its entirety (or possibly use the six-subtest GPAB) and should refrain from selectively administering alternative batteries. We acknowledge that this recommendation runs counter to current neuropsychological practice and "multibattery" approaches to assessment, both of which appropriate subtests from a variety of sources to construct idiosyncratic batteries intended to test clinical hypotheses and address specific referral needs. Our position is a conservative one, recognizing that multibattery approaches represent a departure from the standardized administration procedures used to develop test norms. An alternative would be to use brief ability tests that were designed for short administration and that have known reliability and validity when used in this manner (e.g., Glutting, Adams, & Sheslow, 2000; Psychological Corporation, 1999). Practitioners must be cautious about trading away the advantages inherent in following a standardized protocol in exchange for a briefer, more flexible, and allegedly more "focused" battery of unknown validity.

FUTURE DIRECTIONS: THE DEVELOPMENT OF SB5

As this chapter is being written, preliminary item tryouts are beginning for the development of the SB5, which is planned to be released in Spring 2003. There obviously may be substantial changes between the proposed version of the test and the final published edition, with actual data playing a substantial role in the translation from theory to the published incarnation. Even so, the theory and planning behind the SB5 deserve some comment.

The plans for the SB5 seek to honor the Binet tradition while also incorporating current methodology and theories of intelligence (J. Wasserman, personal communication, February 17, 2000). One major change is explicit adoption of the multi-level model of intelligence expounded by Cattell, Horn (see Horn & Noll, 1997), and Carroll (see Carroll, 1993). The goal in developing the SB5 is to include items that will adequately sample all eight hypothesized specific ability factors: fluid reasoning, general knowledge, quantitative reasoning, working memory (previously short-term memory), long-term memory, auditory processing, visual–spatial ability, and processing speed. The battery is also expected to include measures of procedural knowledge in an effort to measure Gc, or crystallized ability. If the data support the desired model, then the plan would be for SB5 to yield factor scores for each of these specific abilities. In a departure from tradition, the SB5 will probably express these standard scores in a metric with $M = 100$ and $SD = 15$ (not the $SD = 16$ of previous Stanford–Binet scales). The expectation is that the SB5 will also yield three superordinate scores: Verbal Ability, Nonverbal Ability, and a full-scale Composite score reflecting the single best estimate of psychometric *g* obtained from the test. Each of these constructs will also have an observed scaled score that practitioners will calculate as part of the standard scoring of the battery.

Current plans also include other features designed to make the test more appealing to clinicians. One is to utilize a balanced set of verbal and nonverbal indicators for each of the eight specific ability factors, addressing a historical criticism of the SB instruments as overemphasizing verbal abilities. A second feature is the plan to generate linking samples of youths completing the SB5 and either the Wechsler Individual Achievement Test—Second Edition or the Achievement tests from the WJ-III. This would substantially facilitate the analysis of IQ–achievement discrepancies when these popular measures of academic achievement are used. Perhaps most notable of all, the SB5 is expected to extend through adulthood, with

new norms reaching ages 80–90. Extensive validity studies are also planned, comparing the SB5 with a variety of other measures of cognitive ability, as well as looking at performance on the SB5 within special populations (defined using independent research and diagnostic criteria). This would be an important contribution to the body of knowledge, in addition to being useful data for test interpretation, because such an approach would avoid the circular reasoning that plagues much research in this area. Too often researchers have used a test to define a diagnosis (e.g., mental retardation or learning disabilities), and then demonstrated that this group shows different performance on other measures of the same construct—without acknowledging the tautology of this approach (Glutting, McDermott, Watkins, Kush, & Konold, 1997).

Also under consideration is a return to the age scale format used in versions prior to SB4. This would eliminate individual subtest scores from the SB5 and make the factor scores the lowest level of analysis. This approach would be consistent with the goal of making SB5 developmentally sensitive, allowing a blending of items designed to measure the same construct across different ages, without requiring a formal change in subtest. Item-level factor analysis (or analysis of parcels developed using IRT) would guide the organization of items as indicators of the specific ability factors.

This return to an age scale format is likely to be controversial, given the amount of clinical lore surrounding the practice of subtest interpretation. However, this change is also consistent with the best evidence currently available, which shows that subtest interpretation is fraught with psychometric problems (Macmann & Barnett, 1997; McDermott et al., 1990, 1992) and generally has failed to deliver the promised improvements in interpretation, diagnosis, or intervention (Watkins & Kush, 1994). Because of their greater reliability and larger amount of variance attributable to an underlying cognitive ability (i.e., greater validity), factor scores are more likely to enable clinicians to make finer-grained analyses than simple interpretation of a global score. The planned format for the SB5 could do much to promote good clinical practice in this regard. Excluding the subtests would certainly discourage the scientifically unwarranted practice of interpreting them. At the same time, providing better measures of the specific ability constructs of the Cattell–Horn–Carroll model would equip practitioners to measure distinct cognitive abilities underlying g. It would still be necessary to demonstrate the treatment validity of the different factor scores (cf. Glutting, Youngstrom, et al., 1997; Youngstrom et al., 1999), but these scores would inherently possess better construct validity than subtest scores. We hope that the finished product for the SB5 achieves the goals its developers have set for this revision.

RECOMMENDATIONS

On the basis of this review, we offer the following general recommendations affecting use of the SB4. As with any recommendation or clinical practice, these are subject to change in light of new research findings.

1. Reasonable construct validity is present for the Composite, and the Composite also has accumulated the most evidence for external and criterion-related validity. This is a score that psychologists can interpret on the basis of psychometric principles, empirical evidence, and best practices.

2. SB4 area scores are problematic because of the continued controversy about SB4's factor structure, as well as the current lack of any data showing incremental validity of the area scores surpassing the interpretive value of the Composite. We have also advanced the position that current disagreement about the adequacy of a four-factor model may not necessarily represent a failure of SB4 per se. Nevertheless, until optimal methodological procedures are applied and empirical evidence supports four underlying factors, psychologists would do well to avoid comparing or interpreting these scores.

3. Subtest interpretation should be deemphasized or avoided, on both psychometric and scientific grounds. Subtest interpretation increases the possibility of Type I errors and complicates the assessment process. Most importantly, subtest analysis has yet to demonstrate incremental validity or treatment validity with the SB4 or other major tests of ability.

4. We believe we have amply demon-

strated the hazards psychologists face in constructing their own SB4 short forms. Cases could be made either for administering the test in its entirety, or for using one of the established and validated short forms. The two most documented and empirically supported short forms currently appear to be the four-subtest form (especially as a screener) and the six-subtest GPAB. Better documentation of the effects of subtest sequencing, as well as the establishment of short forms' the external validity, should remain high priorities on the research agenda. Though less glamorous than some investigations, this work would have important applications in an era focusing on efficiency and cost containment in the provision of psychological assessment.

5. SB4 should not be administered to preschoolers believed to have mental retardation. Because of floor effects, the test shows little capacity for detecting moderate to severe retardation at these age levels. Moreover, the WPPSI-R generally supports floors equal to or slightly lower than those of SB4.

6. SB4 provides a sufficient ceiling for the identification of examinees who may be gifted at any age. The breadth of constructs measured and its extended age range also increase the likelihood that SB4 will become a favored instrument for the assessment of giftedness (Laurent et al., 1992). However, it is worth noting that the revisions of the Wechsler scales published after the SB4 also extended their norms to 3.67 or 4 standard deviations (i.e., maximum standard scores of 155 to 160), essentially establishing parity with the SB4 in this respect.

7. We have argued that IQ tests should not be marketed without adequate evidence that they reflect the aptitudes of individuals with disabilities. However, we cannot reasonably hold SB4 to standards that have never been imposed on any IQ tests at the time of their publication. The SB4 clearly met the standard of practice in test development when it was published. It is this standard of practice itself that needs improvement. Currently marketed tests have yet to do an adequate job of documenting the appropriateness of the instrument for individuals with disabilities or other specific populations. In practical terms, the SB4 appears comparable to the other best tests available in technical adequacy in this area.

8. It is critical that examiners control test pieces when evaluating young children (especially the Bead Memory pieces, due to the potential choking hazard).

9. Examiners should inquire about color-blindness or family history of color-blindness, as well as remaining alert to this possibility in their clinical observations during testing. The prevalence of color-blindness is high enough that clinicians will encounter this issue with some frequency, and it can influence performance on some subtests of SB4.

CONCLUSION

At the beginning of this chapter, we stated that no test is entirely without fault or virtue. Perhaps SB4's greatest limitation is that it tries too hard to offer everything psychologists want in an IQ test. Nevertheless, SB4's potential for meeting the avowed purposes of IQ tests is great, and, as is far too rare in the field of test development, the positive features of this instrument outweigh its limitations.

ACKNOWLEDGMENTS

Special thanks to John Wasserman, project coordinator for the development of the Stanford–Binet Intelligence Scale: Fifth Edition (SB5), for discussing the planned revisions while the SB5 project was still in progress. Thanks also to Carla Kmett Danielson, Shoshana Kahana, and Erin McMullen for their help in tracking down references for the various tables.

NOTES

1. The problem of small test pieces extends beyond SB4. Several other tests administered to young children, including the Bayley Scales of Infant Development (Bayley, 1969), contain item pieces so small that they are dangerous. Of course, test publishers could argue that it is the responsibility of psychologists to exercise due caution with test materials. Such a position, however, ignores the likelihood that the publisher will be named in any lawsuit stemming from accidents with test materials. Superseding any financial considerations, it is in the best interest of *children* that test materials be safe.
2. Although the BTBC was replaced recently by

the Boehm Test of Basic Concepts—Revised (Boehm, 1986) the original BTBC (Boehm, 1971) was used so that current results would be comparable to those reported by Kaufman (1978).

3. Some of the best evidence for four-factor solutions relies on data using median subtest correlations collapsed across age ranges (e.g., Keith et al., 1988; Thorndike, 1990). Two considerations argue for caution in interpreting these solutions: (a) Using median correlations may hide developmental change (Sattler, 1992); and (b) such approaches have ignored the problem of missing data. Vastly different numbers of participants completed subtests within each age group. Tables B.1 to B.17 in the *Technical Manual* report the "pairwise" *n*'s for each correlation, and numbers can fluctuate dramatically (e.g., *n*'s from 38 to 314 for age 12; see Table B.11) within a given age group. These sampling problems are likely to contribute to technical difficulties in estimating factor structure, and they bias observed results in unknown ways.
4. The abbreviated batteries discussed earlier do not suffer from this problem, because they are composed of subtests 1 through 6 in SB4's administration sequence.

REFERENCES

Aiken, L. R. (2000). *Psychological testing and assessment.* Boston: Allyn & Bacon.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). New York: Macmillan.

Atkinson, L. (1991). Short forms of the Stanford–Binet Intelligence Scale, Fourth Edition, for children with low intelligence. *Journal of School Psychology, 29,* 177–181.

Bayley, N. (1969). *Bayley Scales of Infant Development: Birth to two years.* New York: Psychological Corporation.

Binet, A., & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectual des anormaux. *L'Année Psychologique, 11,* 191–244.

Boehm, A. E. (1971). *Boehm Test of Basic Concepts: Manual.* New York: Psychological Corporation.

Boehm, A. E. (1986). *Boehm Test of Basic Concepts—Revised: Manual.* New York: Psychological Corporation.

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

Boyle, G. J. (1990). Stanford–Binet IV Intelligence Scale: Is its structure supported by LISREL congeneric factor analyses? *Personality and Individual Differences, 11,* 1175–1181.

Bracken, B. A. (1985). A critical review of the Kaufman Assessment Battery for Children (K-ABC). *School Psychology Review, 14,* 21–36.

Bradley-Johnson, S. (2001). Cognitive assessment for the youngest children: A critical review of tests. *Journal of Psychoeducational Assessment, 19,* 19–44.

Carpentieri, S. C., & Morgan, S. B. (1994). Brief report: A comparison of patterns of cognitive functioning of autistic and nonautistic retarded children on the Stanford–Binet—Fourth Edition. *Journal of Autism and Developmental Disorders, 24,* 215–223.

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

Caruso, J. C. (2001). Reliable component analysis of the Stanford–Binet Fourth Edition for 2- to 6-year-olds. *Psychological Assessment, 13,* 261–266.

Carvajal, H. H., & Gerber, J. (1987). 1986 Stanford–Binet abbreviated forms. *Psychological Reports, 61,* 285–286.

Carvajal, H. 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. 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. H., Hayes, J. E., Lackey, K. L., & Rathke, M. L. (1993). Correlations between scores on the Wechsler Intelligence Scale for Children—III and the General Purpose Abbreviated Battery of the Stanford–Binet IV. *Psychological Reports, 72,* 1167–1170.

Carvajal, H. H., McVey, S., Sellers, T., & Weyand, K. (1987). Relationships between scores on the General Purpose Abbreviated Battery of Stanford–Binet IV, Peabody Picture Vocabulary Test—Revised, Columbia Mental Maturity Scale, and Goodenough–Harris Drawing Test. *Psychological Record, 37,* 127–130.

Carvajal, H. H., Parks, J. P., Bays, K. J., & Logan, R. A. (1991). Relationships between scores on Wechsler Preschool and Primary Scale of Intelligence—Revised and Stanford–Binet IV. *Psychological Reports, 69,* 23–26.

Carvajal, H. 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. (1940). A culture-free intelligence test, I. *Journal of Educational Psychology, 31,* 161–179.

Choi, H.-S., & Proctor, T. B. (1994). Error-prone subtests and error types in the administration of the Stanford–Binet Intelligence Scale: Fourth Edition. *Journal of Psychoeducational Assessment, 12,* 165–171.

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,* 112–130.

Coren, S., Ward, L. M., & Enns, J. T. (1999). *Sensation and perception.* New York: Harcourt Brace Jovanovich.

Cronbach, L. J. (1987). *Review of the Stanford–Binet*

Intelligence Scale: Fourth Edition (1007–310). Lincoln, NE: Buros Institute of Mental Measurements.

DeLamatre, J. E., & Hollinger, C. L. (1990). Utility of the Stanford–Binet IV abbreviated form for placing exceptional children. *Psychological Reports, 67,* 973–974.

Delaney, E. A., & Hopkins, T. F. (1987). *Examiner's handbook: An expanded guide for Fourth Edition users.* Chicago: Riverside.

Elliott, C. D. (1990). *Differential Ability Scales: Introductory and technical handbook.* San Antonio, TX: Psychological Corporation.

Embretson, S. E. (1999). Issues in the measurement of cognitive abilities. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 1–16). Mahwah, NJ: Erlbaum.

Flanagan, D. P., & Alfonso, V. C. (1995). A critical review of the technical characteristics of new and recently revised intelligence tests for preschool children. *Journal of Psychoeducational Assessment, 13,* 66–90.

Flynn, J. R. (1984). IQ gains and the Binet decrements. *Journal of Educational Measurement, 21,* 283–290.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54,* 5–20.

Glutting, J. J., Adams, W., & Sheslow, D. (2000). *Wide Range Intelligence Test manual.* Wilmington, DE: Wide Range.

Glutting, J. J., & Kaplan, D. (1990). Stanford–Binet Intelligence Scale, Fourth Edition: Making the case for reasonable interpretations. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence and achievement* (pp. 277–295). New York: Guilford Press.

Glutting, J. J., McDermott, P. A., & Stanley, J. C. (1987). Resolving differences among methods of establishing confidence limits for test scores. *Educational and Psychological Measurement, 47,* 607–614.

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

Glutting, J. J., Youngstrom, E. A., Oakland, T., & Watkins, M. (1996). Situational specificity and generality of test behaviors for samples of normal and referred children. *School Psychology Review, 25,* 94–107.

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

Greene, A. C., Sapp, G. L., & Chissom, B. (1990). Validation of the Stanford–Binet Intelligence Scale: Fourth Edition with exceptional black male students. *Psychology in the Schools, 27,* 35–41.

Gridley, B. E., & McIntosh, D. E. (1991). Confirmatory factor analysis of the Stanford–Binet: Fourth Edition for a normal sample. *Journal of School Psychology, 29,* 237–248.

Grossman, J. J. (Ed.). (1983). *Classification in mental retardation.* Washington, DC: American Association on Mental Deficiency.

Grunau, R. E., Whitfield, M. F., & Petrie, J. (2000). Predicting IQ of biologically "at risk" children from age 3 to school entry: Sensitivity and specificity of the Stanford–Binet Intelligence Scale IV. *Journal of Developmental and Behavioral Pediatrics, 21,* 401–407.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory.* New York: Wiley.

Harris, S. L., Handleman, J. S., & Burton, J. L. (1990). The Stanford–Binet profiles of young children with autism. *Special Services in the Schools, 6,* 135–143.

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.

Hendershott, J. L., Searight, H. R., Hatfield, J. L., & Rogers, B. J. (1990). Correlations between the Stanford–Binet, Fourth Edition and the Kaufman Assessment Battery for Children for a preschool sample. *Perceptual and Motor Skills, 71,* 819–825.

Hernandez, D. J. (1997). Child development and the social demography of childhood. *Child Development, 68,* 149–169.

Hopkins, T. F. (1988). Commentary: The Fourth Edition of the Stanford–Binet: Alfred Binet would be proud. . . . *Measurement and Evaluation in Counseling and Development, 21,* 40–41.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179–185.

Horn, J. L. (1968). Organization of abilities and the development of intelligence. *Psychological Review, 79,* 242–259.

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.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53–91). New York: Guilford Press.

Husband, T. H., & Hayden, D. C. (1996). Effects of the addition of color to assessment instruments. *Journal of Psychoeducational Assessment, 14,* 147–151.

Jaeger, R. M. (1984). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology, 57,* 253–270.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Kamphaus, R. W. (1993). *Clinical assessment of children's intelligence.* Boston: Allyn & Bacon.

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

Kaufman, A. S. (1978). The importance of basic con-

cepts in the individual assessment of preschool children. *Journal of School Psychology, 16,* 207–211.

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. (1988). Confirmatory factor analysis of the Stanford–Binet Fourth Edition: Testing the theory–test match. *Journal of School Psychology, 26,* 253–274.

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

Kelley, T. L. (1927). *Interpretation of educational measurements.* Yonkers, NY: World Books.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

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.

Kline, R. B., Snyder, J., Guilmette, S., & Castellanos, M. (1992). Relative usefulness of elevation, variability, and shape information from WISC-R, K-ABC, and Fourth Edition Stanford–Binet profiles in predicting achievement. *Psychological Assessment, 4,* 426–432.

Kline, R. B., Snyder, J., Guilmette, S., & Castellanos, M. (1993). External validity of the profile variability index for the K-ABC, Stanford–Binet, and WISC-R: Another cul-de-sac. *Journal of Learning Disabilities, 26,* 557–567.

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. (1999). Stability of the SB:FE and K-ABC for young children from low-income families: A 5-year longitudinal study. *Journal of School Psychology, 37,* 315–332.

Kyle, J. M., & Robertson, C. M. T. (1994). Evaluation of three abbreviated forms of the Stanford–Binet Intelligence Scale: Fourth Edition. *Canadian Journal of School Psychology, 10,* 147–154.

Laurent, J., Swerdlik, M., & Ryburn, M. (1992). Review of validity research on the Stanford–Binet Intelligence Scale: Fourth Edition. *Psychological Assessment, 4,* 102–112.

Lavin, C. (1995). Clinical applications of the Stanford–Binet Intelligence Scale: Fourth Edition to reading instruction of children with learning disabilities. *Psychology in the Schools, 32,* 255–263.

Lavin, C. (1996). The Wechsler Intelligence Scale for Children—Third Edition and the Stanford–Binet Intelligence Scale: Fourth Edition: A preliminary study of validity. *Psychological Reports, 78,* 491–496.

Levy, P. (1968). Short-form tests: A methodological review. *Psychological Bulletin, 69,* 410–416.

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

Macmann, G. M., & Barnett, D. W. (1997). Myth of the master detective: Reliability of interpretations for Kaufman's "intelligent testing" approach to the WISC-III. *School Psychology Quarterly, 12,* 197–234.

McCallum, R. S., & Karnes, F. A. (1990). Use of a brief form of the Stanford–Binet Intelligence Scale (Fourth) for gifted children. *Journal of School Psychology, 28,* 279–283.

McCallum, R. S., Karnes, F. A., & Crowell, M. (1988). Factor structure of the Stanford–Binet Intelligence Scale (4th Ed.) for gifted children. *Contemporary Educational Psychology, 13,* 331–338.

McCallum, R. S., & Whitaker, D. P. (2000). The assessment of preschool children with the Stanford–Binet Intelligence Scale Fourth Edition. In B. A. Bracken (Ed.), *The psychoeducational assessment of preschool children* (3rd ed., pp. 76–102). New York: Grune & Stratton.

McCarthy, D. A. (1972). *Manual for the McCarthy Scale of Children's Abilities.* New York: Psychological Corporation.

McCormick, R. L. (1956). A criticism of studies comparing item-weighting methods. *Journal of Applied Psychology, 40,* 343–344.

McCrowell, K. L., & Nagle, R. J. (1994). Comparability of the WPPSI-R and the S-B:IV among preschool children. *Journal of Psychoeducational Assessment, 12,* 126–134.

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

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

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Moreland, K. L., Eisman, E. J., & Dies, R. R. (1998). *Benefits and costs of psychological assessment in health care delivery: Report of the Board of Professional Affairs Psychological Assessment Workgroup, Part I.* Washington, DC: American Psychological Association.

Moffitt, T. E., & Silva, P. A. (1987). WISC-R Verbal and Performance IQ discrepancy in an unselected cohort: Clinical significance and longitudinal stability. *Journal of Consulting and Clinical Psychology, 55,* 768–774.

Molfese, V., Yaple, K., Helwig, S., & Harris, L. (1992). Stanford–Binet Intelligence Scale (Fourth Edition): Factor structure and verbal subscale scores for three-year-olds. *Journal of Psychoeducational Assessment, 10,* 47–58.

Nagle, R. J., & Bell, N. L. (1993). Validation of Stanford–Binet Intelligence Scale: Fourth Edition abbreviated batteries with college students. *Psychology in the Schools, 30,* 227–231.

Nagle, R. J., & Bell, N. L. (1995). Validation of an item-reduction short form of the Stanford–Binet Intelligence Scale: Fourth Edition with college students. *Journal of Clinical Psychology, 51,* 63–70.

Naglieri, J. A. (1988a). Interpreting area score variation on the fourth edition of the Stanford–Binet Scale of Intelligence. *Journal of Clinical Child Psychology, 17,* 225–228.

Naglieri, J. A. (1988b). Interpreting the subtest profile

on the fourth edition of the Stanford–Binet Scale of Intelligence. *Journal of Clinical Child Psychology, 17,* 62–65.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77–101.

Olridge, O. A., & Allison, E. E. (1968). Review of Wechsler Preschool and Primary Scale of Intelligence. *Journal of Educational Measurement, 5,* 347–348.

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.

Prewett, P. N. (1992). Short forms of the Stanford–Binet Intelligence Scale: Fourth Edition. *Journal of Psychoeducational Assessment, 10,* 257–264.

Psychological Corporation. (1997). *Wechsler Adult Intelligence Scale—Third Edition, Wechsler Memory Scale—Third Edition technical manual.* San Antonio, TX: Author.

Psychological Corporation. (1999). *Wechsler Abbreviated Scale of Intelligence manual.* San Antonio, TX: Author.

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.

Rosenthal, B. L., & Kamphaus, R. W. (1988). Interpretive tables for test scatter on the Stanford–Binet Intelligence Scale: Fourth Edition. *Journal of Psychoeducational Assessment, 6,* 359–370.

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.

Rust, J. O., & Lindstrom, A. (1996). Concurrent validity of the WISC-III and Stanford–Binet IV. *Psychological Reports, 79,* 618–620.

Saklofske, D. H., Schwean, V. L., Yackulic, R. A., & Quinn, D. (1994). WISC-III and SB:FE performance of children with attention deficit hyperactivity disorder. *Canadian Journal of School Psychology, 10,* 167–171.

Sattler, J. (1992). *Assessment of children* (3rd ed.). San Diego, CA: Author.

Saylor, C. F., Boyce, G. C., Peagler, S. M., & Callahan, S. A. (2000). Brief report: Cautions against using the Stanford–Binet-IV to classify high-risk preschoolers. *Journal of Pediatric Psychology, 25,* 179–183.

Schwean, V. L., Saklofske, D. H., Yackulic, R. A., & Quinn, D. (1993). WISC-III performance of ADHD children. *Journal of Psychoeducational Assessment (WISC-III Monograph),* pp. 56–70.

Silverstein, A. B. (1993). Type I, Type II, and other types of errors in pattern analysis. *Psychological Assessment, 5,* 72–74.

Simpson, M., Carone, D. A., Burns, W. J., Seidman, T., Montgomery, D., & Sellers, A. (2002). Assessing giftedness with the WISC-III and the SB4. *Psychology in the Schools, 39,* 515–524.

Smith, D. K., & Bauer, J. J. (1989). *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, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12,* 102–111.

Spearman, C. (1923). *The nature of intelligence and the principles of cognition.* London: Macmillan.

Spruill, J. (1988). Two types of tables for use with the Stanford–Binet Intelligence Scale: Fourth Edition. *Journal of Psychoeducational Assessment, 6,* 78–86.

Spruill, J. (1991). A comparison of the Wechsler Adult Intelligence Scale—Revised with the Stanford–Binet Intelligence Scale (4th Edition) for mentally retarded adults. *Psychological Assessment, 3,* 133–135.

Thorndike, R. L. (1973). *Stanford–Binet Intelligence Scale: Form L-M, 1972 norms tables.* Boston: Houghton Mifflin.

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

Thorndike, R. L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin.

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

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

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986c). *Technical manual, Stanford–Binet Intelligence Scale: Fourth Edition.* Chicago: Riverside.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41,* 321–327.

Vernon, P. E. (1950). *The structure of human abilities.* London: Methuen.

Vernon, P. E. (1987). The demise of the Stanford–Binet Scale. *Canadian Psychology, 28,* 251–258.

Volker, M. A., Guarnaccia, V., & Scardapane, J. R. (1999). Short forms of the Stanford–Binet Intelligence Scale: Fourth Edition for screening potentially gifted preschoolers. *Journal of Psychoeducational Assessment, 17,* 226–235.

Wainer, H. (1999). The most dangerous profession: A note on nonsampling error. *Psychological Methods, 4,* 250–256.

Watkins, M. W. (1996). Diagnostic utility of the WISC-III Developmental Index as a predictor of learning disabilities. *Journal of Learning Disabilities, 29,* 305–312.

Watkins, M. W., & Kush, J. C. (1994). Wechsler subtest analysis: The right way, the wrong way, or no way? *School Psychology Review, 23,* 640–651.

Watkins, M. W., Kush, J. C., & Glutting, J. J. (1997). Prevalence and diagnostic utility of the WISC-III SCAD profile among children with disabilities. *School Psychology Quarterly, 12,* 235–248.

Wechsler, D. (1967). *Manual for the Wechsler Preschool and Primary Scale of Intelligence.* New York: Psychological Corporation.

Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intelligence—Revised: Manual.* San Antonio, TX: Psychological Corporation.

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

Weiner, E. A., & Stewart, B. J. (1984). *Assessing individuals.* Boston: Little, Brown.

Wersh, J., & Thomas, M. R. (1990). The Stanford–Binet Intelligence Scale—Fourth Edition: Observations, comments and concerns. *Canadian Psychology, 31,* 190–193.

Woodcock, R. W., & Johnson, M. B. (1990a). *Woodcock–Johnson Psychoeducational Battery—Revised Edition.* Allen, TX: DLM Teaching Resources.

Woodcock, R. W., & Johnson, M. B. (1990b). *Woodcock–Johnson Psychoeducational Battery—Revised Edition: Examiner's manual.* Allen, TX: DLM Teaching Resources.

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

Ysseldyke, J. E., & Stevens, L. J. (1986) Specific learning deficits: The learning disabled. In R. T. Brown & c. R. Reynolds (Eds.), *Psychological perspectives on childhood exceptionality: A handbook* (pp. 381–422). New York: Wiley.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432–442.