Behavior Evaluation Scale–2 (BES-2)

The Behavior Evaluation Scale–2 (S. B. McCarney & Leigh, 1990) is a 76-item rating scale developed for students enrolled in kindergarten through grade 12. The manual provides a brief description of the rationale for and importance of assessing emotional/behavioral disorders. The manual also provides clear administration, scoring, and interpretation guidelines, as well as technical information about the standardization process. Five scale scores are provided on the BES-2: Learning Problems, Interpersonal Difficulties, Inappropriate Behavior, Unhappiness/Depression, and Physical Symptoms/Fears. Educators who have had at least one month of interaction with a student may complete the BES-2 rating scale. Estimated time to complete the rating form is 15 to 20 minutes. The authors suggest six primary uses of this scale:

1. Screen for behavior problems
2. Assess the behavior of students who are referred
3. Assist in the diagnosis of behavior disorders or emotional disturbance
4. Contribute to the development of individual educational programs
5. Document progress resulting from behavioral interventions
6. Collect data for research purposes

All items on the BES-2 are stated in observable and measurable terms. In addition, great effort has been made to link behavioral descriptions to federal guidelines established by Public Law 94–142, as well as to Bower's (1981) widely used definition of
emotionally disturbed/behaviorally disordered children. To reduce the subjective nature of teachers' ratings, a 7-point scale is provided: 1, never or not observed; 2, less than once a week; 3, approximately once a month; 4, approximately once a week; 5, more than once a week; 6, daily at various times; 7, continuously throughout the day. The authors claim that rating the frequency of behaviors can reduce the need for more expensive methods of direct and continuous observation by school personnel.

Scores

Five types of scores can be obtained from the BES-2: weighted scores of individual items, subscale raw scores, subscale standard scores, a quotient representing performance on the total scale, and percentile ranks for each subscale and total scale. Scale scores are computed by transferring ratings for each of the 76 items to a data-summary sheet on the rating protocol. Each subscale lists the specific item numbers that compose the five individual scales. Item ratings are multiplied by a preestablished weighting factor and summed by scale. The sums of the weighted scores for each of the five subscales are converted to standard scores (mean = 10; standard deviation = 3) and percentile ranks. A total quotient score (mean = 100; standard deviation = 15) and percentile rank are computed by summing the five subscale standard scores and using appropriate conversion tables provided in the appendix.

Norms
The BES-2 was standardized on a completely new standardization sample. A total of 2,272 students from 31 states were selected to represent four geographic regions of the United States. A sample of 568 general education teachers administered the BES-2 to randomly selected students from their classes. The number of participating children was distributed fairly equally across the K–12 grade levels (range = 137–207). Demographic characteristics of the sample provide a good match to the 1980 U.S. census data for gender, race, ethnicity, geographic area, and educational status of parents. A notable exception appears to be an underrepresentation of Hispanic American and Asian American students within the standardization sample.

Reliability

Two types of reliability data are offered for the BES-2: internal consistency and test–retest. Internal-consistency coefficients ranged from .75 to .95, with 20 of 24 calculated coefficients exceeding a coefficient alpha of .88. Total-scale coefficients were either .97 or .98 across all grade levels. Stability of scores obtained on typically developing children (N = 82) from 10 states and on children with emotional or behavior disorders (N = 108) from 13 states was assessed at an interval of 10–14 days. Spearman correlation coefficients between two sets of obtained scores (test and retest) exceed or round to .90 (range = .89–.97). All correlations were found to be statistically significant at the .001 level of confidence. No evidence of interrater agreement is provided.

Validity
Studies were conducted to provide evidence of validity for the BES-2. Three types of validity are presented: content validity, criterion-related validity, and construct validity. Appropriateness of scale items (content validity) was initially evaluated by a large sample of classroom teachers and special education personnel with expertise in the area of behavioral disorders. In addition, during the revision process, a sample of 675 professionals from 31 states reviewed each item to determine its appropriateness for describing educational behavioral problems. All items on the final version were judged by 95 percent of the experts as appropriate descriptors of problem behavior.

To evaluate concurrent criterion-related validity, the BES-2 was compared with the Teacher Rating scale of the Behavior Rating Profile (BRP; L. Brown & Hammill, 1978). Twenty-six students from elementary and secondary grades, identified as behaviorally disordered and receiving special education services in Missouri, were assessed on both measures. Five of six scores on the BES-2 produced coefficients that were statistically significant at the .01 and .05 confidence levels (range = .44–.81). One subscale, Learning Problems, did not correlate at all with the BRP ($r = .01$). In another study, teachers' professional judgments were compared with BES-2 results. A total of 190 teachers rated students from general ($N = 82$) and special education ($N = 108$) classrooms (K–12) on the BES-2 after having answered the question, "Relative to other students of this age, how would you generally rate this student's classroom behavior?" A 9-point scale was used by the teachers to rate students. Behavioral descriptions were provided for 1, "Student has significant behavior problems in comparison to others"; 5, "Student's behavior is about the same as that of others"; and 9, "Student has excellent classroom..."
behavior in comparison with others." No additional guidelines on how to rate students on the 9-point scale were provided to teachers. The five BES-2 subscales and total score produced significant correlation coefficients at the .01 level of confidence except for students with emotional or behavior disorders rated on Physical Symptoms/Fears. Correlations ranged from .45 to .75 for general education students and from .15 to .59 for students with behavior disorders.

Evidence for construct validity is based on studies showing that students identified as behaviorally disordered exhibited significantly more psychopathology across all five subscales.

**Summary**

The BES-2 is a 76-item rating scale for children enrolled in kindergarten through grade 12. Three subscales produce raw scores that are converted to standard scores (mean = 10; standard deviation = 3). The total score is the only score that can be represented as a percentile rank. The standardization procedures appear to be appropriate (norm tables are provided by age–gender groups), although reliability measures were not reported in this manner. Validity studies support the criterion-related and construct validity of the BES-2 as a device that is adequate for screening purposes.