Critique of the Behavioral and Emotional Rating Scale- Second Edition (BERS-2)

Chapman University

Critique of the Behavioral and Emotional Rating Scale- Second Edition (BERS-2)

Description of the test

The Behavioral and Emotional Rating Scale: A Strength-Based Approach to Assessment-Second Edition (BERS-2) is a standardized, norm-referenced scale written by Michael H. Epstein and published by PRO-ED Inc. in 2004. It is a strength-based assessment designed to identify a child's behavioral and emotional strengths rather than their deficits. The BERS-2 is a multifaceted assessment that gathers data from the child, the parent or caregiver, and the teacher or therapist. The assessment retrieves data from each source through rating scales and open-ended questions to find detailed strengths. It contains three separate rating scales designed for each of the three participating sources: the parent, the teacher, and the child him/herself.

Content and Use

The BERS-2 includes a Parent Rating Scale (PRS), Teacher Rating Scale (TRS), Youth Rating Scale (YRS), a Summary Form, and a manual for the administrator's use. It is an untimed assessment but on average takes about ten minutes per rating scale. Each rating scale is organized as a separate form which is useful to the administrator because it allows them to keep responses separate and organized and also allows multiple scales to be given at once. For example, the teacher and the parent can work on their own scales at the same time since they are separate. This is beneficial for time management. Each scale consists of five core subscales including Interpersonal Strength, Family Involvement, Intrapersonal Strength, School Functioning, and Affective Strength. The Interpersonal Strength subscale measures a child's ability to control their emotions and behaviors. The Family Involvement subscale measures their relationship with their family. The Intrapersonal Strength subscale measures a child's sense of accomplishment and competence. The School Functioning subscale measures a child's

competence in classroom tasks. Finally, the Affective Strength subscale measures a child's ability to express feelings to others and accept affection. The PRS and YRS contain 57 items that aim to identify a child's strengths based on the subscales detailed above. The TRS contains 52 items for the same purpose. Each question is posed as a Likert scale with four anchors (0-3). The participant is to rate the student's status in the past three months on each question on a scale of 0-3. The criteria for the rating scale are as follows:

3 = very much like the student or very much like you

2 = like the student or like you

1 = not much like the student or not much like you

0 = not at all like the student or not at all like you

Additionally, the PRS and YRS include an optional Career Strength subscale meant to measure the child's ability for career development. All of the rating scales also include eight open-ended questions regarding the child's favorite hobbies, sports, school subjects, friends, and other questions aimed at detailing strengths and preferences. The anchors are clear and precise making it easy for raters to pick the option that applies best. Subjective words such as "often" are avoided making the options clear. However, veracity scales such as questions involving the words "always" or "never" are not included allowing for social desirability to occur. Additionally, there is no reverse scoring which typically protects a rater from scoring the same number to finish faster. The questions are all similarly asked allowing raters to answer quickly and sometimes without much thought.

The BERS-2 manual also details the process of rater selection. Each rater for the PRS and TRS should be very familiar with the child being assessed. Raters of the TRS are typically

teachers, counselors, or school psychologists. The PRS rater is typically the parent, guardian, foster parent, or house parent. Children using the YRS should be at least 11 years old.

Furthermore, the BERS-2 manual explicitly states examiner qualifications. It states that examiners should understand the foundation of the instrument, understand the construction and statistical characteristics, understand general principles of norm-referenced assessments, be able to administer the BERS-2 itself, and have knowledge in interpreting results. Professionals who may have the necessary training for this assessment include teachers, psychologists, and social workers. The manual also recommends an examiner to have practiced administering and scoring the assessment at least three times before administering an official one.

Once an examiner is qualified to use the BERS-2, the test can be used for many purposes. First, it is useful in identifying children with limited behavioral and emotional strengths. The tool is not developed to diagnose students with emotional or behavioral disorders, but it can indicate those at risk. Practitioners can also use the BERS-2 for identifying strengths and weaknesses for intervention given that it is a strengths-based assessment. Additionally, the assessment is useful in developing IEP goals and treatment plans, and documenting progress on those goals. It is especially useful for periodic evaluations of students receiving professional services. Finally, the BERS-2 is useful for data collection for research because it quantifies children's strengths using standard scores.

Overall, the manual is user friendly for those with enough knowledge in the subject manner. It is neatly organized with a table of contents and includes an ample amount of information on norm-referenced samples, reliability, validity, and test administration and scoring. The manual even provides visual examples of completed rating scales to assist in scoring. The language is

easy to understand, but laypersons who are not trained in statistical terms may not fully comprehend the material. However, it is simple language for those with some training.

Standardization Sample and Norms

The BERS-2 includes normative data for each rating scale. It provides two sets of data for the TRS where one set is based on children not identified with an emotional or behavioral disorder (NEBD) and the other set is based on children with an emotional or behavioral disorder (EBD). One data set is provided for each the PRS and YRS.

All data sets, except the YRS, include data for children ages 5-18 years. The YRS does not include data for children under 11-years-old because children younger than that are not supposed to rate themselves. Nonetheless, the NEBD TRS data set and the YRS contain an adequate sample size per age group. The majority of the age groups reach the minimum sample size of 100, and a couple reach 200. Some age groups, such as 16-year-olds, are not adequately represented because they do not have enough participants of that age. The sample size per age group for the PRS and TRS EBD are concerning. While all ages between 5 and 18 are represented, the majority of age groups do not meet the minimum requirement of 100 participants. This is concerning because most students are not adequately represented by the normative data for the EBD TRS and the PRS.

While the age groups are not as representative as they should be for some of the rating scales, all data sets are geographically represented. Samples include participants from 31-34 states that stretch around the country. States from the west coast, east coast, south, and Midwest are included. Moreover, the educational attainment of parents for all of the data sets match the percentage of the U.S. population quite well. For example, the majority of the population for both the normative sample and the overall U.S. do not have a bachelor's degree, and less than

10% have a master's, professional, or doctoral degree. These aspects of the data samples represent the national population adequately. Additionally, the BERS-2 does not mention SES, but instead only mentions family income in dollars. SES is generally fuzzy because we do not know how people were identified as being low, middle, or high class. However, family income is simple to identify because there is little room for interpretation and confusion when it comes to hard numbers. The normative data sets generally represent the U.S. distribution of family income. For example, the percentage of the PRA and YRA samples match the U.S. percentage accurately. However, in the TRA data sets, the percentage of participants with a family income of 75,000 and over is almost 10% less than the U.S. population percentage for that same group. This means that students with a family income of over 75,000 are underrepresented in the TRA data sets.

Moreover, the overall sample sizes of each data set are representative of both males and females. Each sample size consists of about half and half. When looking at further details of both sexes, users can see that most of the characteristics are represented equally in both males and females as well. For instance, sample sizes of age groups among males and females are quite similar and so forth. In addition, race is also included for the normative samples, but the only identifications are white, black, and other. "Other" is not specified. Whether a participant is Hispanic or not is included separately. Throughout all the data sets, the white and black population is represented accurately. The "other" population and Hispanic population is underrepresented in both of the TRS samples, but equally represented for the PRS and YRS. Additionally, the disability status of the NEBD TRS, PRS, and YRS samples are accurately representative of the disability status of the overall U.S. population. Disabilities identified are learning disabilities, speech-language impairment, mental retardation, and other disabilities. No

disability is of course included as well. Since the EBD TRS sample population are all identified with an emotional or behavioral disability, it is rightfully not representative of the U.S. population.

Overall, the BERS-2 normative data sets are representative of the U.S. population in terms of sex, race, income, educational attainment, and disability status. The most concerning areas are the age groups because they are not all equally or adequately represented. Given that there are 4 different normative data sets, users must be aware of the normative samples for each rating scale and know which to use. This can be overwhelming when trying to decide whether the student being assessed is adequately represented in the normative data because they might be under the PRS, but not the TRS and so forth. The information given by the BERS-2 is useful, but examiners should always have the manual to look at details regarding the samples for each rating scale.

Scores and Interpretation

The BERS-2 includes raw scores, percentile ranks, scaled scores, and a strength index. Percentile ranks are included because they are easily understood; however, the manual does a great job at warning users of the common misunderstandings of percentile ranks. For example, it states that percentile ranks are not interval data and do not measure distance, therefore, they do not represent equal differences in the measured attributes and behaviors. While it is great that the manual explains this, users need to take caution when sharing the percentile ranks with families and other interested parties because small differences can be exaggerated. Moreover, scaled scores and the strength index are both a type of standard score with a mean of 100 and standard deviation of 15. Given that the BERS-2 is measuring behavior, standard scores should not be

used because behaviors are not normally distributed. T-scores are the best option for behavioral measures, but this assessment does not include them.

Scoring Procedures

The BERS-2 manual includes visual examples showing how to compute the total raw score for each subscale. After adding all of the numbers in each column representing each subscale, a raw score is computed for each. This step is completed for each rating scale: the YRS, PRS, and TRS. On the front page, each rating scale packet includes a row for each subscale where the raw score, percentile rank, and scaled score is inputted. The Appendix in the manual includes normative tables for examiners to convert the raw scores into percentile ranks and scaled scores. However, before converting the scores, the examiner must decide whether they are comparing the student to the NEBD or EBD norms because there are separate tables for each. However, computing scores based on both normative tables might be beneficial to compare the student to both their NEBD and EBD peers. Since both tables are available, users should take advantage and compare their students to both populations to ensure best practice.

Each normative table is also separated by male and female. The fact that there are multiple normative tables is impressive because normative scores should not be treated as a "one size fits all." The option of multiple tables allows the examiner to compare the student's scores where they fit best (i.e. male or female, EBD or NEBD). One single normative table forces all test-takers to fit into the same box which is hardly ever the case.

Additionally, the Strength Index is found for each rating scale. To find the strength index, the five subscale scaled scores are added together for each rating scale. Using the found sum, examiners then look to the appendix to convert the sum to the strength index. A separate percentile rank corresponding to the strength index is also included. A strength of the BERS-2 is

that a strength index is found for each rating scale instead of combining each of the sources' data into one score. The different raters can have different perspectives and experiences with the student; therefore, comparing scores contrived from each rating scale accounts for those differences. If there was one single score, we run the risk of losing the importance of those different perspectives. By having three different strength indexes to look at, interpreters can see if strengths vary among settings (school, home, etc.) or differ between sources. For example, if the strength index from the TRS is concerning, but is not in the PRS or YRS, then it would push interpreters to seek more information regarding what is going on at school. Perhaps the student just does not have a good relationship with the teacher or classmates or there is another underlying educational problem. By separating scores by rating scale, the BERS-2 supports a multi-faceted approach to serving children because the source's data are not all squished into one number.

Interpretation

The BERS-2 manual reminds users that scores alone are not diagnostic and that other evaluative measures such as record review, observations, interviews and other tests are necessary. This mention is very impressive given that many test makers aim to sell their tests as a magical diagnostic measure. Hopefully, users read and understand that the score received from the BERS-2 is only part of what is needed to diagnose.

Moving forward, once norms are determined and scores are computed, the user must utilize their findings to determine the likelihood of EBD. In the BERS-2, low scores are concerning or considered deviant because it is measuring emotional and behavioral strengths. In other words, a low score indicates a lack of strength in a certain area. The manual also includes a range of scores and what those scores mean when compared to other peers (average, above

average, etc.). With all this information, the BERS-2 makes it quite simple for users to determine where a student falls and whether they are likely to have EBD or not. It is great that the manual refers to low scores as a likelihood of having EBD rather than just having EBD. This might remind users that a single score does not diagnose but is only additional data.

Overall, the manual is honest about the importance of scores found by the BERS-2. However, users should understand that standard scores may not accurately represent the behaviors measured in this assessment. The use of T-scores would make the BERS-2 even more impressive. Additionally, the BERS-2 instructs users how to calculate confidence intervals but does not explain which should be used or why. Without knowledge of the confidence interval, users do not know how much error is accounted for. If a 68% confidence interval is used, the scores contain a lot more error versus if a 95% confidence interval was used. The confidence interval is critical to determine how much error surrounds the score. Users cannot interpret a score's meaning with any certainty without knowing the confidence interval. The BERS-2 should better explain and utilize this concept.

Psychometric Properties

Reliability

The BERS-2 details three types of reliability evidence: internal consistency, time sampling, and interrater reliability. Internal consistency measures the consistency in response patterns. To measure this, Cronbach's alpha was calculated at different age intervals for the TRS NEBD and EBD samples, PRS, and YRS. An average coefficient for each subscale and the strength index is computed and shown in a table. Most of the coefficient averages for the TRS NEBD sample are above .90 which is adequate for educational decisions and can be considered a strength of the test. Even though this level of reliability is impressive, 10% of the method error is

still unknown. Furthermore, most of the TRS EBD, PRS, and YRS average coefficients reach .80 but not .90. This level of reliability is adequate for research purposes and screeners but not for making educational decisions which is supposed to be a purpose of the BERS-2. This is concerning because users are most likely using the assessment to make decisions regarding their students even though most of the rating scales are not adequately reliable for this purpose. Furthermore, the BERS-2 includes Cronbach's alpha coefficients for certain subgroups such as males, females, white students, black students, Hispanic students, and emotionally disturbed students. The test maker realizes that just because an assessment is considered reliable for a general population, that does not mean it is equally reliable for subgroups. The idea to include coefficients for these subgroups is a strength, but not all of them exhibit adequate levels. Users need to consider the coefficients for each subgroup and each rating scale to determine how reliable their score is. Some coefficients are appropriate for educational purposes and meet the .90 guideline, but others do not even reach the .80 guideline for research purposes. These should be interpreted with caution. Sointu, Savolainen, Lambert, Lappalainen, and Epstein (2014) administered the BERS-2 in Finland and also yielded Cronbach's alpha coefficients above .80. The BERS-2 proves to be cross-culturally consistent and reliable in other countries, but again decisions based on these scores should be made cautiously because the coefficients are only adequate for research purposes.

Additionally, the BERS-2 considers time sampling through test-retest. Six studies were conducted to determine the stability of the BERS-2 results over time. Only two of the six studies waited six months to retest. The other four studies retested after two weeks. A behavioral or emotional disorder is not expected to vanish after a short period of time; therefore, researchers should have waited more than two weeks to retest and determine whether another factor was at

play. Nonetheless, coefficients for the short-term, two-week studies met the appropriate strength guidelines of .80 and above. This is expected because results should remain stable after such a short amount of time. More importantly, the results for the long-term study (6 months) meet the strength guidelines with coefficients of .60 or above. This is considered a strength because after six months, the assessment appears to remain relatively stable with at least 36% of the temporal error accounted for. However, each of the six studies has a sample size of less than 100 which is not representative. The results of the studies seem impressive, but the small sample size reduces the reliability. Moreover, there is no alternative test for retesting purposes. Especially when there is a short time period between retesting, raters may remember how they answered previously and do the same to get done faster or in hopes of yielding similar results. Without alternative tests, carryover effects may come into play and reduce temporal reliability.

Lastly, the BERS-2 tackles interrater reliability. Interrater reliability refers to the amount of error in examiner variability in rating individuals. The manual cautions that interrater observations of behavior usually yield lower correlation coefficients because of the difference in demands placed on children in different settings. Three studies were conducted to test interrater reliability between teacher/teacher, teacher/parent, and parent/student. The teacher/teacher study yielded coefficients above .80 which is expected because they are rating within the same setting. Guidelines expect the coefficients to be between .60-.80 and the results were impressively higher than that. In addition, the teacher/parent study yielded correlation coefficients between .54-.67, with a .20 outlier. The majority of the coefficients surpass the guidelines indicating that more source error is accounted for. The final study between parents and students all yielded coefficients of .50 and above surpassing the .40 guideline. The BERS-2 portrays a strength in

interrater reliability because the correlation coefficients surpass general guidelines and accounts for more source error than expected.

Epstein (2004) argues that the BERS-2 is impressively reliable, but that is not the whole truth. At a glance, the numbers look impressive and promising, but when looking at the details, there are areas of concern. However, the manual does a good job at detailing reliability expectations and where the BERS-2 stands in comparison. Overall, the BERS-2 shows promising evidence proving its reliability, but users should be aware of its downfalls when interpreting scores.

Validity

The BERS-2 manual provides evidence for content-description validity, criterion-prediction validity, and construct-identification validity.

Content-Description Validity. Content-description validity examines if a test basically measures what it is supposed to measure. This was originally examined in the original BERS by first conducting a literature review of all research available on strength-based assessment, resilience, developmental psychopathology, and protective factors. Extensive research on existing assessments that measure affective characteristics was also conducted. Then, data were collected from an expert panel consisting of professionals from education, mental health, child welfare, and other social service fields to find out what behaviors were seen as strengths in children. However, no information is given on the expert panel, so users do not know their qualifications. For example, graduate students would not be considered experts compared to those with doctorate's degrees and professors. Qualifications of the expert panel are necessary to ensure true validity. Nonetheless, Epstein gathered responses from these professionals, sorted through them, and sent out multiple additional surveys to narrow down the items until he had

around 127 items left. To narrow down those items even further, item discrimination and factor analyses were performed. To establish item discrimination, a study was conducted that ensured items were present in children with an emotional or behavioral disorder but absent in children without EBD. A factor analysis was also conducted to determine how the items correlate together. Those that did not covary were removed which left the BERS with 68 items.

Content validity was already established in the BERS by the measures described above; however, the PRS, YRS, and the Career Strength Subscale were added into the BERS-2 and needed validity verification. Similar procedures from the BERS item identification were implemented for the Career Strength items. The same developed items from the TRS were used in the PRS and YRS.

Additionally, for the BERS-2, an item analysis was conducted on all the norm samples from the NEBD and EBD TRS, PRS, and YRS to ensure item discrimination. The manual states that because the medians of the item coefficients were above .35 across all subscales, content validity was proven. The BERS-2 also contains a differential item functioning analysis to determine biased items. Differential item functioning occurs when people from different racial groups perform differently on items even if they have the same level of emotional and behavioral strength. This analysis specifically compared Black students versus others and Hispanic students versus others. Results yielded that less than 5% of the comparisons were significant indicating that the test may contain a small amount of bias regarding race or ethnicity. While it is great for the BERS-2 to include differential item functioning and admit that some bias exists, there are other minority groups aside from Blacks and Hispanics and there is no data supporting their performance. Items may be even more biased towards unmentioned groups, but users have no information for comparison.

Overall, if the expert panel that was chosen to verify adequate items were indeed qualified, the BERS-2 appears to show content validity and measure what it is supposed to. The BERS-2 has a good amount of data supporting this through multiple item analyses and item discrimination.

Criterion-Prediction Validity. Criterion-prediction validity checks test performance by comparing results to similar measures or other criteria such as achievement, grades, and diagnoses. The BERS-2 informs users that "criterion-prediction" validity is used instead of "criterion-related" validity, but they mean the same thing. This is a little confusing because it may cause users to assume that predictive measures were conducted when the BERS-2 actually used concurrent measures. The manual states that concurrent validity, where one test is given immediately after the other, is more appropriate. Nonetheless, the BERS-2 provides studies on the TRS, PRS, and YRS to prove criterion-related validity.

Six studies were conducted regarding the TRS with the purpose of finding a correlation between the BERS-2 and other assessments measuring similar skills and behaviors. Since the BERS-2 is a strength-based assessment, behavioral measures were expected to negatively correlate with BERS-2 strengths, and other strength measures were expected to positively correlate with BERS-2 strengths. Of the six studies for the TRS, all of the correlations were as expected. Based on Hopkins (2002) correlation magnitude standards, all correlations between measures were moderate to large (.3-.7). Similarly, two studies were conducted to measure the correlation between other measures and the BERS-2 for the PRS. All correlations were as expected and were considered large by Hopkins standards (.5-.7). Two additional studies were conducted similarly for the YRS. Correlations fell in the moderate to large range and were as expected in terms of positive and negative relationships.

In critiquing the studies that prove criterion-prediction validity, the first issue that pops out is the correlation magnitude standards. The manual considers .5-.7 as large based on Hopkins standards; however, .7 or above is typically considered large. By these standards, most of the correlations between the BERS-2 and similar measures are only moderate, and some are even considered low. Additionally, no other criteria are accounted for. For example, the studies do not mention correlations between the BERS-2 and achievement, grades, or diagnoses. A measure cannot be considered truly valid if all related criteria are not accounted for. While the studies prove moderate correlations between the BERS-2 and other similar measures, additional correlations between similar criteria are necessary.

Construct-Identification Validity. Construct validity measures the degree to which underlying traits of a test can be identified and to which the concepts of a test match the theoretical model it is based on. The manual states several constructs that the BERS-2 is thought to reflect. For example, underlying constructs include the ability to differentiate children with EBD and without EBD, the expectation that subscales correlate to an extent, and the expectation that items from each subscale be highly correlated with the total score of the subscale. To prove construct validity based on these concepts, the manual reflects on the norm samples from each subscale. Males, females, Whites, Blacks, Hispanics, and children with EBD were separated into subgroups and their scores were compared. All of the subgroups' scores landed in the average range, except for the EBD group which landed in the below average range. This was expected because it proves that the BERS-2 can identify those with emotional and behavioral problems and those without them. Utilizing these subgroups is beneficial because it appropriately compares children who have these EBD related issues to those that do not. The subgroups differentiated by ethnicity and gender are also beneficial in establishing nonbiased measures

between populations. Uhing, Mooney, and Ryser (2005) conducted a separate study measuring validity in the BERS-2 PRS and YRS. They compared scores of children who had EBD and children who did not have EBD. In both rating scales, results yielded statistically significant differences between children with EBD and children without EBD. Specifically, children who had EBD received lower scores on the BERS-2. This evidence further supports construct validity because the BERS-2 appears to appropriately differentiate children with emotional and behavioral strength deficits from their typical peers.

Additionally, the subscale scores are correlated in the .37-.87 range with a mean of .65. The mean correlation coefficient establishes a moderate correlation even though the manual regards this as large based on Hopkins standards. The manual also states that correlations between subscales should not be too high because that would indicate that they are not unique from each other. Based on this idea, the mean correlation coefficient appears adequate as it is moderately correlated but does not challenge the uniqueness of the subscales. A factor analysis was also conducted to prove that the traits of the BERS-2 support the theoretical model in which it is based. Four indexes were computed where each index met its own criteria to be considered "a good fit" and serves as evidence that the BERS-2 is well established and supports its theoretical model.

Overall, the BERS-2 shows promising evidence supporting its validity. The content and construct validity evidence appear to be more adequate than the evidence for criterion validity due to the moderate correlations and lack of comparison to other important criteria. Nonetheless, users can feel comfortable that the BERS-2 is supported with an ample amount of data regarding validity.

Conclusions

In conclusion, the BERS-2 is an easy to follow and overall impressive assessment tool in identifying children's emotional and behavioral strengths. It is great that the manual stresses not to diagnose based on the yielded scores because the test aims to assist in determining the likelihood of EBD. The manual is very honest in what the scores indicate and how professionals should use them. Additionally, it is very clear, organized, and provides visual examples to simplify the administering and scoring process. It is very simple to follow and understand. The BERS-2 also does a good job in including a representative norming sample and appears to be culturally sensitive in terms of reliability. The BERS-2 is advertised as an impressive tool regarding reliability and validity. The data supporting these are good, but not great. The BERS-2 does appear to be consistent and measure what it is supposed to measure; however, some of the statistics are not as strong as advertised. Overall, users can feel confident in this measure, but should nonetheless interpret scores with caution and ensure that the student they are testing is adequately represented.

References

Epstein, M. H. (2004). *Behavioral and Emotional Rating Scale: A strength-based approach to assessment: examiners manual* (2nd ed.). Austin, TX: Pro-Ed.

Hopkins, W.G. (2002). A scale of magnitudes for effect statistics. *A new view of statistics*.

Sointu, E. T., Savolainen, H., Lambert, M. C., Lappalainen, K., & Epstein, M. H. (2014). Behavioral and emotional strength-based assessment of Finnish elementary students: Psychometrics of the BERS-2. *European Journal of Psychology of Education*, *29*(1), 1–19.

Uhing, B. M., Mooney, P., & Ryser, G. R. (2005). Differences in strength assessment scores for youth with and without ED across the Youth and Parent Rating Scales of the BERS-2. *Journal of Emotional & Behavioral Disorders*, *13*(3), 181–187.