The Screening Test for Emotional Problems – Teacher-Report Version (STEP-T): Studies of Reliability and Validity

Measurement and Evaluation in Counseling and Development
2015, Vol. 48(2) 152–160

DOI: 10.1177/0748175614563315
mec.sagepub.com
SAGE

Bradley T. Erford[1], Caitlin Butler[1], and Elizabeth Peacock[1]

Abstract

The Screening Test for Emotional Problems – Teacher Version (STEP-T) was designed to identify students aged 7–17 years with wide-ranging emotional disturbances. Coefficients alpha and test-retest reliability were adequate for all subscales except Anxiety. The hypothesized five-factor model fit the data very well and external aspects of validity were excellent.

Keywords

behavioral assessment, confirmatory factor analysis, school counseling, counseling

An increasing number of children in the United States are identified with emotional disturbances (Rudy & Levinson, 2008), and outcomes for these students are generally poor. Students with emotional disturbances often display patterns of behavior that lead to difficulty achieving success throughout their lives. Rather than taking the traditional "wait to fail approach" (Nelson, Epstein, Griffith, & Hopper, 2007), the most effective way for these students to make progress is through early identification to receive the appropriate academic and behavioral interventions from early on in school (Glover & Albers, 2007). Students with emotional disturbances who are not identified early are at a higher risk for lasting social maladjustment.

Only 1% of children receive special education services for emotional disturbances, while it was estimated that at least 7% of the school-aged population had some type of emotional disturbance. In order to qualify for special education services, students must meet the characteristics identified in the definition established in the Individuals With Disabilities Education Improvement Act (IDEA; U.S. Department of Education, 2004).

Emotional disturbance is defined under the IDEA (U.S. Department of Education, 2004) as: (a) an inability to learn that cannot be explained by intellectual, sensory, or health factors; (b) an inability to build or maintain satisfactory interpersonal relationships with peers and teachers; (c) inappropriate types of behavior or feelings under normal circumstances; (d) a general pervasive mood of unhappiness or depression; or (e) a tendency to develop physical symptoms or fears associated with personal or school problems (Section 300.7c). This definition is vague, leading to numerous interpretations of mental and emotional disorders subsumed therein (Erford, 2013).

While several instruments have been used in schools to identify emotional disturbances,

[1]Educational Specialties Department in the School of Education at Loyola University Maryland, Timonium, MD, USA

Corresponding Author:
Bradley T. Erford, Education Specialties Department in the School of Education at Loyola University Maryland, Timonium Graduate Center, 2034 Greenspring Drive, Timonium, MD 21093, USA.
Email: berford@loyola.edu

none are specifically aligned with the public law definition. Along with the lack of time- and cost-effective methods of assessment, this has made the identification of students at-risk for emotional disturbance particularly difficult. Another aspect contributing to this problem is a lack of confidence that educators often have in administering and interpreting the more complex instruments correctly (Rudy & Levinson, 2008).

Educators often use a combination of instruments in order to assess the criteria laid out in the federal definition. Some of these assessments may include: the Achenbach System of Empirically Based Assessment (ASEBA; Achenbach & Rescorla, 2001), the Beck Depression Inventory – Second Edition (BDI-II; Beck, Steer & Brown, 1996), the Behavior Assessment System for Children – Second Edition (BASC-2; C. R. Reynolds & Kamphaus, 2006), the Conners 3 (Conners, 2008), the Revised Children's Manifest Anxiety Scale – 2nd edition (RCMAS-2; C. R. Reynolds & Richmond, 2008), the Reynolds Adolescent Depression Scale – Second Edition (RADS-2; W. M. Reynolds, 2002), or the Walker Problem Behavior Identification Checklist (WPBIC; Walker, 1983).

Another problem with using these instruments is that not all of them gather information from different informants. Only the ASEBA (Achenbach & Rescorla, 2001), BASC-2 (C. R. Reynolds & Kamphaus, 2006), and Conners 3 (Conners, 2008) assess self-report, teachers, and parents, while the other instruments are self-report only. But the most important weakness is that the items across these instruments vary, making direct comparisons difficult because different respondents are asked different question sets. In screening for emotional disturbance, it is important to obtain and converge the perspectives of caregivers and teachers, due to the differences in the behaviors that can be observed in different environments (Rudy & Levinson, 2008). In addition, research has demonstrated that when parents and teachers share the same problem definition, the outcomes for students are more positive (Nelson et al., 2007). Collecting data from teachers and parents on the same instrument and then comparing the data should yield the most accurate information about a student. This will in turn lead to the ability to make better decisions about appropriate interventions, program placement, and referrals to outside agencies for these students (Erford, Balcom, & Moore-Thomas, 2007). In general, teachers have more opportunities to view children interacting socially, which is an important component that contributes to the accurate measurement of emotional disturbances.

The early identification of students with emotional disturbance is critical to determining the best course of action to take. Educators need a comprehensive, time- and cost-effective method of screening for emotional disturbances that assesses all five of the criteria established in the IDEA definition to help insure that students with this disability are best served by the public schools in the United States.

Purpose of This Study

This article provides results of preliminary studies of reliability and validity of teacher responses to the Screening Test for Emotional Problems – Teacher Version (STEP-T), which can be administered to teachers in order to screen children and adolescents for emotional disturbances as defined by IDEA. It should be noted that self-report (Erford, Short, & Freeman, 2012) and parent report versions (Erford et al., 2007; Erford & Alsamadi, 2012) of the STEP have also been developed. The only difference between the parent, teacher, and student versions of the STEP was the use of the words "home" and "school" in several items, and that the stems for each item relate to the respondent's relationship to the student (e.g., self-report: I tell lies; parent report: My child tells lies; teacher report: The student tells lies). The student (Erford et al., 2012) and parent (Erford & Alsamadi, 2012) versions yielded adequate score reliability and validity for a screening level test.

In order to determine reliability of teacher responses to the STEP, internal consistency and test-retest reliability were explored.

External score validity also was explored through correlating scores on the STEP-T with other tests that also measure emotional problems and behavior disorders, including the Disruptive Behavior Rating Scale – 2nd edition (Erford & Clark, 2011), and Conners 3 – Teacher Short Form (Conners, 2008). Structural validity of STEP-T scores was demonstrated through a study of confirmatory factor analysis (CFA) to determine the fit of the data to the hypothesized five-factor model.

Method

Three independent samples of convenience were used in the following three studies. Participants were volunteers recruited as samples of convenience from several Mid-Atlantic states. No monetary or privilege incentives were issued for participation. Guidelines for approval of a human participants study were followed and parental informed consent and participant assent obtained. Minor-status participants were recruited from public and private schools. Importantly, even though female teachers outnumbered male teachers, no evidence existed that different gendered teachers rate students differentially. The construct-based validity model (Messick, 1995), specifically internal structural aspects and external aspects of validity, were used to study score validity of the STEP-T scores in the following three studies.

Participants

Study 1. Participants were the 415 teachers (70 males, 345 females; 95% White) of 678 students (283 boys, 395 girls) ages 7–17 years ($M = 11.61$, $SD = 2.72$). Of the students, about 61% were White, 21% African American, 11% Hispanic American, 3% Asian American, and 3% "Other." About 21% of students were from urban settings, 74% from suburban settings and 5% were from rural settings. Of the parents of the students, 12% of the fathers did not complete high school, 44% had only a high school diploma, 20% completed some college and 24% were college graduates. Of the mothers, 5% did not complete high school, 41% had only a high school diploma, 26% completed some college, and 28% were college graduates.

Study 2. Participants were the 45 teachers (6 males, 39 females; 96% White) of 85 students (47 boys, 38 girls) aged 7–17 years ($M = 12.18$, $SD = 3.10$) from four schools in the Mid-Atlantic region of the United States. About 73% of the students were White, 12% were African American, 9% were Hispanic American, 4% were Asian American, and 2% "Other." About 89% of students were from urban or suburban settings (communities larger than 2,500 people) and the remaining 11% were from rural settings. Of the parents of the students, 6% of the fathers did not complete high school, 44% had only a high school diploma, 20% had completed some college and 30% were college graduates. About 5% of the mothers did not complete high school, 51% had only a high school diploma, 24% completed some college, and 20% were college graduates.

Study 3. Participants were the 34 teachers (4 males, 30 females; 94% White) of 68 students (41 boys, 27 girls) ages 7–17 years ($M = 11.82$, $SD = 2.85$). Of the students, 74% were White, 16% were African American, 8% were Hispanic American, and 2% "Other." Of the parents of the students, 9% of the fathers did not complete high school, 50% completed a high school diploma, 18% completed some college and 23% were college graduates. About 6% of the students had mothers who did not complete high school, 54% had only a high school diploma, 21% completed some college, and 19% were college graduates.

Instruments

The STEP-T. The STEP-T (Erford et al. 2007) can be individually or group administered to teachers of students 7–17 years of age and scored in less than 15 minutes. The STEP-T consists of 40 items, with eight items assigned to each of five subscales designed to align with the five IDEA categories for identification of emotional disturbances: (a) Academic

Problems (e.g., 16. The student is currently failing at least one major subject area (e.g., reading, math, language, spelling, social studies, science), (b) Behavior Problems (e.g., 18. The student does not obey home or school rules), (c) Social Problems (e.g., 17. The student has few friends), (d) Depression (e.g., 19. The student cries for no apparent reason, or as an inappropriate response to personal or school situations), and (e) Anxiety (e.g., 30. The student worries about school, home, or personal situations). Items were designed to apply to students across the age range of 7–17 years, which spans middle childhood and adolescence. Participants respond to each item on a 3-point scale (i.e., 0 = *do* not *agree*; 1 = *somewhat agree*; and 2 = *very much agree the statement is a problem for the student*). Raw scores for each subscale are determined by simple sum of score procedures and can be converted into *T* scores ($M = 50$, $SD = 10$) and percentile ranks. Coefficients α on a clinical sample were reported to be: Academic Problems (α = .88), Social Problems (α = .76), Behavior Problems (α = .73), Anxiety (α = .79), and Depression (α = .75) (Erford et al., 2007). External (convergent and discriminant) and internal aspects of validity will be presented in the Results section.

Conners 3 – Teacher Short Version (Conners 3-T-S). The Conners 3-T-S is widely used to assess problem behaviors in children and adolescents (Conners, 2008). It is part of a multi-instrument system consisting of both long and short versions for parents, teachers, and self-report. The short form version of the teacher instrument used in this study consisted of 39 items, assessing problematic behaviors across five factors. Each item is rated on a four-point scale: 0 = *not true at all (never, seldom)*, 1 = *just a little true (occasionally)*, 2 = *pretty much true (often, quite a bit)*, 3 = *very much true (very often, very frequent)*. The Conners 3-T-S is appropriate for use with teachers of children ages 6–18 years.

The normative sample for the Conners 3-T-S included teachers of 1,200 students ages 6–18 years: 50 males and 50 females at each age level 6–16 years and a combination age group of 17- and 18-year-olds. Participants' race/ethnicity was approximately representative of the U.S. population. Parental educational level indicated that about 4% of the norm group had parents who did not complete high school, 18% had only a high school diploma, 29% completed some college or vocational programming, and 41% were college graduates.

Factor analysis of items composing the long version of the Conners 3-T led to the selection of 5 to 6 items for each of five Conners 3-T short version subscales, presented here with accompanying coefficients α: Inattention (α = .95; α range = .94–.95), Hyperactivity/Impulsivity (α = .97; α range = .96–.97), Learning Problems/Executive Function (α = .95; α range = .94–.96), Aggression (α = .95; α range = .95–.96), and Peer Relations (α = .92; α range = .90–.94). Test-retest reliability coefficients of scores included: Inattention (r = .80), Hyperactivity/Impulsivity (r = .82), Learning Problems/Executive Functioning (r = .77), Aggression (r = .80), and Peer Relations (r = .72). Factorial and convergent validity are well documented and discussed at length in the Conners 3 manual (Conners, 2008). For example, the Conners Learning Problems subscale correlated r = .66–.92 with the BASC-2 (C. R. Reynolds & Kamphaus, 2006) Learning Problems scale, and the Conners Conduct Problems subscale correlated .79 with the BASC-2 Conduct Problems scale.

The Disruptive Behavior Rating Scale – 2nd Edition – Teacher (DBRS-II-T). The DBRS-II-T (Erford & Clark, 2011) can be individually or group administered to teachers and scored in less than 15 minutes. The DBRS-II-T consists of 35 items, with seven items assigned to each of five subscales: (a) Inattention (α = .91), (b) Hyperactive/Impulsive (α = .92), (c) Antisocial Conduct (α = .88), (d) Oppositional (α = .92), and (e) Anxiety (α = .77). Each item is rated on a 4-point scale (i.e., 0, 1, 2, or 3) where: "[0] means the child rarely or hardly ever displays the behavior; [1] means the child occasionally displays the behavior; [2] means the child frequently displays the behavior; and [3] means the child displays the

behavior most of the time." Raw scores for each subscale are determined by simple sum of score procedures and can be converted into *T* scores ($M = 50$, $SD = 10$) and percentile ranks. External (criterion-related) validity was reportedly adequate for a screening level test. For example, the DBRS-II-T Antisocial Conduct subscale correlated $r = .69$ with the Conners-3 Aggression scale and the DBRS-II-T Hyperactive-Impulsive subscale correlated .82 with the Conners-3 Hyperactivity-Impulsivity scale (Erford & Clark, 2011).

Procedures

Study 1. Teachers were administered the STEP-T and DBRS-II-T according to standardization specifications and in a counterbalanced order to prevent order effects. Computation of coefficients α for each subscale and confirmatory factor analysis (CFA) using Mplus version 5.21 (Muthen & Muthen, 2007) were conducted. Computation of Pearson correlation coefficients between subscales of the STEP-T and DBRS-II-T were conducted to provide evidence of external validity.

Study 2. The STEP-T was administered to the teachers of the 85 participants, then administered again after exactly 14 days. All protocols were scored according to standardization specifications. Subscale standard scores for each administration were used to compute test-retest Pearson correlation coefficients.

Study 3. The STEP-T and Conners-3 were administered to the teachers of the 68 participants in a counterbalanced order in an attempt to reduce systematic error due to administration sequence. All protocols were administered and scored according to standardized specifications and each test's subscale standard scores were used in computations of Pearson correlation coefficients for the sample.

Results and Discussion

Data were analyzed for normality and outliers, and all variables were within normal limits.

Structural (Factorial) Validity

CFA was conducted on teacher responses for the students in Study 1 ($n = 678$) using Mplus version 5.21 (Muthen & Muthen, 2007) CFA procedures to determine how well the scores fit the hypothesized five-factor model. Mplus for weighted least squares mean variance (WLSMV) estimation was used because the STEP-T yields ordinal scores. The output statistics generated by Mplus included residual (error) variance for each item (see Table 1) and standardized factor correlations (see Table 2). These factor correlations were moderate without being so strong that distinct dimensions could not be justified. Residual variance is the variance observed in each variable not accounted for by the five latent factors in the hypothesized model. The results from Table 1 indicate that most items yielded sufficiently low residual variances.

Mplus indicated a comparative fit index (CFI) of .95, Tucker-Lewis Index (TLI) of .98, root mean square error of approximation (RMSEA) of .074 (90% CI = .071–.077), and weighted root mean square residual (WRMR) of 1.504. Dimitrov (2012) suggested that for adequate model fit, CFI or TLI should be ≥ .95, RMSEA should be ≤ .06, and WRMR should be < 2.00. Thus, the results of this CFA indicate an adequate fit of the data to the five-factor model; CFI, TLI, and WRMR indicate an excellent fit, while RMSEA indicated a marginal fit. It is probable that the inclusion of a behaviorally normal sample of school-aged children led to a more homogeneous data set, thereby leading to a less than optimal fit of the present data to the model. Examination of clinical sample participants with more widely varying scores may enhance model fit.

Reliability

Internal consistency reliability coefficients of rationally derived subscales of the STEP-T item responses for the samples used in Studies 1 and 2 are presented in Table 3 (α range = .69–.92). The 14-day test-retest reliability (Pearson) coefficients of rationally derived subscales of the STEP-T responses in study 2 also are presented in Table 3 (r_{tt} range = .69–

Table 1. Residual (Error) Variance for Each Variable (Item).

Item #	Residual variance
1	.158
2	.463
3	.295
4	.240
5	.321
6	.287
7	.200
8	.316
9	.312
10	.314
11	.203
12	.194
13	.402
14	.336
15	.208
16	.164
17	.305
18	.195
19	.482
20	.993
21	.245
22	.245
23	.258
24	.542
25	.434
26	.194
27	.378
28	.221
29	.199
30	.407
31	.154
32	.265
33	.158
34	.349
35	.250
36	.173
37	.374
38	.406
39	.232
40	.317

.85). Overall, participants' responses to the STEP-T items reached adequate levels of response reliability both in terms of internal consistency and test-retest stability for screening level purposes. The αs and r_{tt}s exceeded a .80 criterion for all subscales except the Anxiety subscale. The Academic Performance subscale performed especially well, with αs ≥ .90 in the two samples.

Table 2. Structure Coefficients Among STEP-T Latent Dimensions.

	F1	F2	F3	F4	F5
F1: Academic Performance	—				
F2: Behavioral Problems	.462	—			
F3: Social Problems	.509	.593	—		
F4: Anxiety	.595	.595	.677	—	
F5: Depression	.463	.478	.538	.667	—

External Aspects of Validity

External aspects of validity of the STEP-T responses involved the calculation of Pearson correlations between the standard scores of subscales of the STEP-T, DBRS-II-T ($n = 678$), and Conners 3-T-S ($n = 68$). Sixteen scale comparisons were expected to yield moderate to high correlations, and for ease of identification these comparisons are bolded in Table 3. The STEP-T displayed moderate to high correlations (i.e., $r \geq .45$) with all 16 predicted subscales (100%) measuring similar constructs, indicating acceptable external validity. The .45 criterion was used successfully by Erford et al. (2007) and Erford and Clark, 2011) in previous studies to yield estimates of convergent and discriminant validity while accounting for coefficients inflated by common methods variance (Erford, 2013). Notably, the STEP-T Behavior Problems subscale correlated .78 and .86 with the DBRS-II-T Oppositional and Antisocial Conduct subscales, respectively. Likewise, the STEP-T Academic Problems subscale correlated .75 with the DBRS-II-T Inattention subscale and .72 and .70 with the Conners 3-T-S Inattention and Learning Problems subscales, respectively.

Limitations, Implications, and Conclusions

These preliminary studies of STEP-T score reliability and validity showed that the instrument

Table 3. Reliability and External Validity Coefficients for Teacher Responses to the STEP-T, Disruptive Behavior Rating Scale (DBRS), Conners 3 Teacher Version – Short Form Subscales From Study 3.

	(1)	(2)	(3)	(4)	(5)	α1	α2	r_{tt}
STEP-T subscales								
(1) Academic Performance	—					.92	.90	.85
(2) Behavior Problems	.47	—				.83	.80	.80
(3) Social Problems	.54	.78	—			.85	.84	.81
(4) Anxiety	.36	.48	.57	—		.73	.69	.69
(5) Depression	.57	.72	.79	.70	—	.83	.80	.80
DBRS-II-T subscales								
Inattention	**.75**	.47	.51	.31	.51			
Oppositional	.50	**.78**	**.74**	.50	.68			
Impulsive-Hyperactive	.49	**.52**	**.49**	.34	.46			
Antisocial Conduct	.44	**.86**	**.70**	.46	.64			
Anxiety	.36	.41	.55	**.71**	**.65**			
Conners 3 – Teacher-Report subscales								
Inattention	**.72**	**.45**	.49	.32	.50			
Hyperactivity/Impulsivity	.46	**.48**	**.45**	.27	.43			
Learning Problems	**.70**	.45	.40	.29	.48			
Aggression	.41	**.75**	**.72**	.43	.59			
Family Relations	.35	.40	.30	.25	.35			

Note. n = 678, except for the Conners 3-T-S coefficients where n = 68. Bolded coefficients indicate predicted external validity; α1 = coefficient α for Study 1; α2 = coefficient α for Study 2; r_{tt} = test-retest reliability for Study 2.

performed as an efficient screening tool for school-aged students. Despite the promising results, there were several limitations of the studies that should be considered. First, in order to adequately compare test results across versions of the test, it must be shown that parents, teachers, and students interpret and respond to items in a similar manner. If items were not considered equivalent, the interpretations of subscale scores offered by different respondents may vary. It was unclear if there was item equivalence for parents, students, and teachers across the three versions of the STEP. Future research should seek to determine item equivalence for all respondents.

Second, in order to enhance factor stability and structural validity, changes should be made to several items, as suggested by CFA study results. Several items could be replaced based on the lack of variance contributed, and several other items with substantial cross-loading should be revised in order to more closely measure the desired item characteristic. However, because the same 40 items were used across all three respondent versions of the STEP, any changes to item configuration or structure modeling must be done in concert across all respondent versions. Third, use of the 3-point scale may lower reliability and validity estimates of low frequency items. Therefore, future research should explore if a 4- or 5- point rating scale would be a more efficient method to increase response heterogeneity and increase the STEP's psychometric integrity even further.

Fourth, the studies' samples were limited based on region, socioeconomic status, and participant self-selection. Therefore, future studies should incorporate more diverse samples. For example, participants from the Mid-Atlantic and Southern U.S. regions were included in these studies, but future research should include participants from other U.S. geographic regions to increase generalizability. Data were not coded for school or individual teacher-respondent, so no secondary analyses are possible. Thus, additional study is needed to ensure the STEP-T is generalizable for use with ethnically diverse youth in urban settings. Lastly, because participant

self-selection is a common problem in social science research, the studies may be somewhat biased. The teacher and student participants were self-selected, which raises the concern that the participants opting not to participate potentially offered different characteristics or perspectives not represented in the studies.

Despite these limitations, the studies were pertinent in providing information on the STEP-T and how it can be used as a screening tool for students with emotional disturbances. Since the current screening and intervention process used in public schools is somewhat unreliable (Cullinan & Sabornie, 2004), future research on the STEP-T may help to provide a consistent, efficient method for screening school-aged students with emotional disturbances to allow for early intervention.

The studies provided evidence that the STEP-T yields reliable and valid scores. It is also cost-efficient and allows for early intervention. Educators, especially teachers, are pressured to dedicate as much time to instruction as possible; therefore, it is imperative to have a quick and efficient way to screen students at-risk for emotional problems. The STEP-T only requires 15 minutes to administer and score, thus providing a user friendly screening option. Additionally, the STEP-T is cost-efficient (i.e., free) and allows for further testing and decision making for students with special needs. A final strength of the STEP-T is the ability to facilitate early intervention for students with emotional disturbances. After the STEP-T is administered, the results could help to provide direction for the implementation of prevention plans at an earlier stage of development. It is imperative that educators have access to screening tools that allow for early intervention for students with emotional disorders, and the STEP-T could play a critical role in early detection and intervention for students with emotional disturbances.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the Achenbach System for Empirically Based Assessment (ASEBA)*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-Second Edition (BDI-II)*. San Antonio, TX: Psychological Corporation.

Conners, C. K. (2008). *Manual for the Conners 3*. North Tonowanda, NY: Multi-Health Systems.

Cullinan, D., & Sabornie, E. (2004). Characteristics of emotional disturbances in middle and high school students. *Journal of Emotional and Behavioral Disorders*, *12*, 157–167. doi:10.1177/10634266040120030301

Dimitrov, M. D. (2012). *Statistical methods for validation of assessment scale data in counseling and related fields*. Alexandria, VA: American Counseling Association.

Erford, B. T. (2013). *Assessment for counselors* (2nd ed.). Belmont, CA: Cengage Wadsworth.

Erford, B. T., & Alsamadi, S. C. (2012). The Screening Test for Emotional Problems–Parent Report (STEP-P): Studies of reliability and validity. *Measurement & Evaluation in Counseling & Development*, *45*, 170-180. doi: 10.1177/0748175611432643

Erford, B. T., Balcom, L., & Moore-Thomas, C. (2007). The Screening Test for Emotional Problems (STEP): Studies of reliability and validity. *Measurement and Evaluation in Counseling and Development*, *39*, 209-225.

Erford, B. T., & Clark, K. (2011). Technical analysis of the Disruptive Behavior Rating Scale – Second Edition – Teacher Version. *Measurement and Evaluation in Counseling and Development*, *44*, 3-15. doi: 10.1177/0748175610391605

Erford, B. T., Short, S., & Freeman, A. (2012). The Screening Test for Emotional Problems – Self-Report (STEP-S): Studies of reliability and validity. *Assessment for Effective Intervention*, *37*, 251-256. doi: 10.1177/1534508411413759

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School*

Psychology, *45*, 117–135. doi:10.1016/j.jsp.2006.05.005

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749. doi:10.1037/0003-066X.50.9.741

Muthen, L. K., & Muthen, B. O. (2007). *Mplus statistical analysis with latent variables user's guide*. Los Angeles, CA: Author.

Nelson, R., Epstein, M., Griffith, A., & Hopper, J. (2007). Factor structure, internal consistency, and interrater reliability of the early childhood behavior problem screening scale. *Topics in Early Childhood Special Education*, *27*, 148–154. doi:10.1177/02711214070270030501

Reynolds, C. R., & Kamphaus, R. W. (2006). *Manual for the Behavior Assessment System for Children–Second Edition (BASC-2)*. Circle Pines, MN: American Guidance Service.

Reynolds, C. R., & Richmond, B. O. (2008). *Manual for the Revised Children's Manifest Anxiety Scale–2nd Edition (RCMAS-2)*. Los Angeles, CA: Western Psychological Services.

Reynolds, W. M. (2002). *Manual for the Reynolds Adolescent Depression Scale–Second Edition (RADS-2)*. Odessa, FL: Psychological Assessment Resources.

Rudy, H. L., & Levinson, E. M. (2008). Best practices in the multidisciplinary assessment of emotional disturbances: A primer for counselors. *Journal of Counseling & Development*, *86*, 494–504. doi:10.1002/j.1556-6678.2008.tb00537.x

U.S. Department of Education. (2004). *The Individuals With Disabilities Education Improvement Act* (P.L. 108-446). Washington, DC: Government Printing Office.

Walker, H. M. (1983). *The Walker Problem Behavior Identification Checklist (WPBIC)* (2nd ed.). Los Angeles, CA: Western Psychological Services.

Author Biographies

Bradley T. Erford, PhD, is a professor in the School Counseling Program of the Education Specialties Department in the School of Education at Loyola University Maryland.

Caitlin Butler is a graduate of the School Counseling Program of the Education Specialties Department in the School of Education at Loyola University Maryland.

Elizabeth Peacock is a graduate of the School Counseling Program of the Education Specialties Department in the School of Education at Loyola University Maryland.