2016

Initial Development and Validation of the Student Wellbeing Teacher-Report Scales

Anthony Joseph Roberson
Louisiana State University and Agricultural and Mechanical College, tonyjroberson@gmail.com

INITIAL DEVELOPMENT AND VALIDATION OF THE STUDENT WELLBEING TEACHER-REPORT SCALES

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Arts

in

The Department of Psychology

by
Anthony J. Roberson
B.S., Truman State University, 2014
December 2016

Dedicated in memory of Sr. Carole (1940–2016) for her commitment to education, appreciation of the importance of mental health, and embrace of life lived well.

Table of Contents

List of Tables

Abstract

Given that youth mental health is associated with their success in school and in life more broadly, it is important that school-based psychological service providers embrace best-practice prevention and intervention strategies that target mental health when working with student populations One line of study in this area has begun exploring the incorporation of a *dual-factor model* of mental health within universal screening systems in schools. The dual-factor model is differentiated from the traditional unidimensional mental health model, which focuses on the presence or absence of psychopathology, by conceptualizing mental health alternatively as consisting of both psychopathology and wellbeing dimensions. The present study involved the preliminary development and validation of the Student Wellbeing Teacher-Report Scales (SWTRS)—a pair of brief behavior rating scales intended to function as screening tools for measuring two indicators of the wellbeing dimension of youths' mental health at school: "feeling good" and "functioning well." Specifically, the study involved drafting pilot items for the SWTRS and explored their latent factor structure, concurrent validity with school-related outcomes (i.e., attendance, academic achievement, and time on-task), as well as concurrent and incremental validity in comparison with psychopathology screeners. Results suggested that the SWTRS items may better represent two context-specific indicators of youths' wellbeing—academic engagement and prosocial behavior—rather than the hypothesized "feeling good" and "functioning well" dimensions. The SWTRS also demonstrated incremental validity and were uniformly stronger predictors of all school-related concurrent outcomes compared to the psychopathology scales. Implications for theory and future research are discussed.

Keywords: youth wellbeing, school mental health, universal screening

Introduction

Universal Screening for Youth Mental Health

The substantive importance of addressing mental health concerns among youth has been well established in a tremendous number of research findings. The National Institute of Mental Health found that among adolescents ages 13–18, more than 46% live with some form of mental illness (Merikangas et al., 2010). Youth mental illness has been linked with many deleterious outcomes, such as adolescents with depression being at higher risk for later substance dependence (Marmorstein, Iacono, & Malone, 2010), more internalizing and externalizing behavior symptoms in youth predicting clinical panic attacks (Mathyssek, Olino, Velhurst, & van Oort, 2012), and the formation of depressive traits in early adolescence predicting depressive episodes later in life (Rudolph & Klein, 2009). Further, lower performance on cognitive, academic achievement, and short-term memory assessments has been predicted by greater severity of anxiety, depression, and social withdrawal symptoms (Rapport, Denney, Chung, & Hustace, 2001). Externalizing behavior issues have also predicted low concurrent academic achievement in school (Bradshaw, Buckley, & Ialongo, 2008). Beyond these individual-level concerns, Keyes (2007) noted that the economic toll of mental health care in the U. S. in 1999 amounted to approximately $160 billion, making it the third costliest health care expense after cardio-vascular disease and physical rehabilitation. Given this evidence that less-than-optimal mental health results in significant negative outcomes for the individual and society at large, developing and implementing systems of care that connect at-risk youth with mental health services should be a priority in order to curb later harmful effects associated with psychological disorder (Lane, Oakes, & Menzies, 2010).

Given that the school environment is most frequently the de facto mental health care provider for young people (Burns et al., 1995), it follows that the goals of both prevention and intervention efforts may best be achieved through enhancing school-based mental health service delivery systems (Strein, Hoagwood, & Cohn, 2003). A prerequisite to delivering quality services to struggling students is determining which students are the best candidates to receive such support; that is, which students are most at-risk for future negative outcomes. Student selection can be accomplished in a variety of ways, but in conventional practice, identification typically begins with a referral from a teacher or parent experiencing concern with the student in class or at home (Strein et al., 2003). The school psychologist, or other school-based mental health professional, then takes on the case and begins the assessment and intervention process to understand and remedy the student's presenting issues. Though this method has been widely employed over the years, alternative student identification models have been gaining traction to address some of the shortcomings of the traditional referral paradigm.

The worth and utility of local population screening procedures to identify students in need of mental health services is receiving increasing acknowledgement from both researchers and practitioners (Albers & Kettler, 2014; Dowdy et al., 2014). Typically administered to populations of students (e.g., all students in a classroom, all students in a school), universal screeners can offer school professionals useful information concerning the mental health functioning of the student body (Albers & Kettler, 2014). Among the many advantages universal screening can offer, the ability to calculate local norms is of great worth to practitioners looking to assess the prevalence and magnitude of specific dimensions of mental health functioning (Dowdy et al., 2010). Once norms are calculated, they can be used in different ways to help school professionals better understand their student population and identify students at-risk for

deleterious outcomes. One of these ways is to compare individual student scores to the rest of the student population. If the individual student shows significantly elevated symptoms relative to the symptom level of their peers, the school psychologist should follow-up with the student to get more information about the presenting problems. This same logic of comparing individual screener scores to some norm extends further to comparing between larger groups of students. A school psychologist may be interested in investigating the prevalence and severity of internalizing and externalizing behavior symptoms in their school and may want to compare these to other similar schools or between classrooms. Following screening, the class- or school-wide norms from other classes or schools can be used to determine if there is a systemic issue influencing the problems (Dowdy et al., 2010). The data may suggest that a general environmental influence is having a global negative effect on students across the school, rather than relatively few students experiencing problems for idiosyncratic reasons.

Perhaps most importantly, researchers and school practitioners are recognizing the importance of screening instruments as a critical component in effectively transforming school-based mental health service delivery from the traditional reactionary model (i.e., students referred for services only after severe problems have manifested) to a paradigm that emphasizes early detection of symptoms to inform prevention efforts (Dowdy et al., 2014). This approach aligns with the goals of incorporating data-based decision making strategies (National Association of School Psychologists, 2010; Armistead & Smallwood, 2014) in service of a stronger public health model (Strein et al., 2003) and establishing multi-tiered systems of support (MTSS) within schools to enhance efficiency of service delivery resource allocation and positive behavioral interventions (Individuals with Disabilities Education Improvement Act, 2004). Moreover, the National Association of School Psychologists recommends MTSS as a best-

practice approach to mental health service delivery in schools (Stoiber, 2014). Typical MTSS models contain three levels of service delivery arranged along a triangular continuum, with all students receiving universal interventions at the bottom tier. Students who continue to show significant difficulty in the domain of interest despite the universal supports are promoted to increasingly specialized and intensive intervention tiers (Stoiber, 2014). Screening instruments play an integral role in the successful implementation of a MTSS scheme within a school system as they can be used as the initial assessment gate for problem identification at the lowest tier (Albers & Kettler, 2014). Once the screening process identifies certain students as having elevated risk, a second gate of follow-up assessments can be employed to further hone in on the students most in need of services at higher tiers.

One of the key areas of research with universal screening in schools is investigating the technical adequacy and applied utility of youth mental health screeners. Literature in this area suggests that screening for mental health may be integral in early detection efforts of deleterious symptoms. For instance, a study conducted with a random sample of 472 elementary school students involved teacher ratings of the students' adaptive and problem behavior frequency using different behavior rating scales for comparison (Kamphaus, DiStefano, Dowdy, Eklund, & Dunn, 2010). Results indicated that scores from the Behavioral and Emotional Screening System (BESS), a 27-item teacher-report screening instrument for rating the frequency of student problem behaviors, significantly correlated with several concurrent outcome measures from an omnibus problem behavior scale and student academic records. Some of these include moderate correlations with math and English language arts (ELA) grades (r = -.45), a strong correlation with the omnibus Behavior Assessment System for Children, Second Edition, Teacher Rating Scales (BASC-2 TRS-C) internalizing problems scale (r = .52), and very strong correlations with

the BASC-2 TRS-C externalizing problems (r = .76), school problems (r = .82), and adaptive skills (r = -.82) scales. Though several other studies of this kind have been conducted supporting the concurrent and predictive validity of scores derived from universal screeners measuring mental health problems (e.g., Eklund & Dowdy, 2013; Lane et al., 2009), scholarship on elementary teacher-report mental health screening is still nascent and warrants further development.

Conceptualizing Youth Mental Health for Universal Screening

Student contact with universal mental health screening and promotion initiatives during the elementary school years is one of the critical school-based strategies for limiting the progression of negative psychological and behavioral symptoms during important developmental years (Lane, Oakes, & Menzies, 2010). However, to reap the benefits of universal screening, it is important for practitioners to utilize high-quality assessment instruments. Glover and Albers (2007) outlined three broad domains to consider when evaluating the quality of a universal screening instrument: (a) appropriateness for the intended use (e.g., do the constructs measured help determine risk level?), (b) technical adequacy (e.g., to what extent are scores derived from the screener reliable and valid?), and (c) usability (e.g., how feasible is it to administer the screener in a real-world context?).

Aspects of the appropriateness consideration have also been discussed in other works, such as the recommendation from Hayes, Nelson, and Jarrett (1987) that assessment processes should inform treatment in a useful way. Lane, Oakes, and Menzies (2010) note that, in addition to other feasibility considerations such as monetary cost, developers of screening instruments should endeavor to strike a balance between including as few items as possible while maintaining strong psychometric properties of the screener to minimize respondent fatigue or the

likelihood that assessment interferes with other school activities. Finally, while creating an instrument with sufficient technical adequacy is a more complex undertaking that cannot be demonstrated in a single study, establishing robust psychometric qualities is no less important than the other considerations. Technical adequacy can be determined by assessing the strength of an instrument's reliability (e.g., internal consistency, inter-rater reliability) and validity dimensions (e.g., predictive/concurrent validity, incremental validity). Developing and validating screening instruments that have all three of the recommended qualities is crucial to progress research on universal mental health screening in schools.

Although much empirical attention has paid to the "technical adequacy" and "usability" dimensions of screening as outlined by Glover and Albers (2007), far less empirical work has focused on the "appropriateness" of mental health indicators being measured and used in school-based screening practice. One of the central issues when deciding how to evaluate youth mental health using screeners is considering which dimensions of the mental health construct should be assessed (Glover & Albers, 2007). Historically, the field of school psychology has placed great emphasis on the amelioration of negative behavioral and psychological symptoms in children, especially those that have a harmful effect on school success (Ysseldyke & Reschly, 2014). This intention is noble, as it aims to limit and ameliorate low academic achievement and problem behaviors. Few would disagree that working to minimize the occurrence and severity of such negative outcomes among students is a benefit to their overall life functioning. Yet, there is some disagreement over whether this approach is the most conceptually sound and useful for promoting mental health, as some have argued that this tradition stems from an ideology that incorrectly views the absence of problems as synonymous with the presence of wellbeing (Seligman & Csikszentmihalyi, 2000; Seligman, 2002).

Though school psychological practice based on this philosophy has been longstanding and indeed done much to improve many students' lives, some posit that incorporating assessment of positive aspects of student functioning is necessary to better understand youth mental health (e.g., Suldo & Shaffer, 2008), especially for purposes of identifying students at greater and lesser risk. Consistent with this idea, the World Health Organization (WHO) has suggested an updated definition of mental health that incorporates this positive behavior lens. The WHO views mental health not as the lack of disease or disorder, but as "a state of well-being in which the individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community" (2004, p. 12). If the larger aim of psychological services is to help youth thrive, rather than simply live without problems, school psychologists should consider incorporating aspects of positive functioning into their assessment and intervention efforts, including universal mental health screening (Furlong, Gilman, & Huebner, 2014; Renshaw et al., 2014).

The importance of positive features of mental health has been explored in multiple lines of research that have used various operational conceptions of wellbeing (WB). For example, Keyes (2006; 2007; Keyes & Annas, 2009) views wellbeing as consisting of two related but distinct components: *hedonic wellbeing* (HWB) and *eudaimonic wellbeing* (EWB). HWB is the frequency and duration of an individual's positive emotional experiences and their overall satisfaction with life; also referred to as "feeling good." On the other hand, EWB is an individual's appraisal of how well they are functioning socially and psychologically; also known as "functioning well." In a study involving 1,234 adolescents, 12–18 years old, Keyes (2006) noted that youth who scored at a high level on at least one indicator of HWB and a high level on over half of the EWB indicators (i.e., mentally healthy or "flourishing") showed fewer symptoms

of depression as compared to those who scored at a low level on at least one symptom of HWB and low levels on more than half of the EWB indicators (i.e., "languishing"). Youth associated with a subjective wellbeing (SWB) profile between "flourishing" and "languishing" were considered "moderately mentally healthy" and showed middling rates of depression symptoms.

Empirical evidence concerning the interaction between aspects of psychopathology (PTH; used broadly here in reference to any negative psychological or behavioral functioning) and WB among youth has suggested that mental health could be conceptualized as a bidimensional rather than unidimensional concept. This *dual-factor model of mental health*—also referred to as *two-continua* or *complete mental health*—suggests that PTH and WB are related but distinct constructs that vary in severity along two dimensions. Greenspoon and Saklofske's (2001) conducted an early exploratory study of dual-factor mental health with a sample of elementary school students. To evaluate dimensions of PTH and SWB among the students, the researchers used several student- and teacher-report subscales that measured aspects of self-concept, interpersonal relationships, personality, temperament, internalizing and externalizing problem behaviors, perceived locus of control, and life satisfaction. Results showed that four distinct mental health categories could reliably be discriminated from each other at rates of 95–300% above chance levels. These categories included: "distressed" —high PTH, low SWB; "externally maladjusted" —high PTH and SWB; "dissatisfied" —low PTH and SWB; and "well adjusted" —low PTH, high SWB. Given that both groups display high levels of PTH, students who would be classified in either the "distressed" or "externally maladjusted" categories are the most likely to be detected in a traditional teacher referral paradigm. However, the authors recognized that there might be important differences in outcomes between the two groups accounted for by the discrepancy in SWB. Further, although students in the "dissatisfied"

category do not display significant symptoms of PTH, they still are lacking in SWB. This group may therefore benefit from an intervention approach aimed at bolstering positive behaviors, rather than problem elimination.

Subsequent research on the dual-factor model of mental health offered support for the hypothesized between-group differences proposed by Greenspoon and Saklofske (2001). For instance, Suldo and Shaffer (2008) used a similar four-group classification system of PTH and SWB with a sample of middle school students to see how well group membership predicted important outcomes in school functioning, social adjustment, and physical health. As predicted, results indicated that students in the "complete mental health" group (i.e., low PTH, high SWB) showed the lowest negative outcome scores (e.g., social problem frequency; school absences) and the highest positive outcome scores (e.g., GPA; motivation and self-regulation). Further, students with elevated levels of PTH showed significantly worse academic performance regardless of SWB level (e.g., GPA for the two high PTH groups was approximately 0.35–0.50 points lower than the "complete mental health" group). These findings were consistent with the traditional view of unidimensional mental health. However, results also showed that students classified as "symptomatic but content" (i.e., high PTH and SWB) endorsed receiving positive support from adults in their lives approximately 17% more frequently than students in the "troubled" group (i.e., high PTH, low SWB), who themselves endorsed experiencing social difficulties at twice the rate of the "symptomatic but content" group. These findings suggest that the presence of higher levels of SWB may function as a buffer from the negative effects of PTH symptoms.

A follow-up study by Suldo, Thalji, and Ferron (2011) investigated the longitudinal predictive value of combined measurement of aspects of student SWB and PTH after one year.

Results again confirmed classic conceptions of mental health. For instance, those with elevated externalizing symptoms at Time 1 were significantly more likely to earn lower GPAs (accounting for 6% of the variance) and have worse school behavior at Time 2 (accounting for 5% of the variance). Furthermore, students with elevated internalizing PTH symptoms, regardless of SWB level, missed on average one more day of school than the low PTH groups. However, support for the dual-factor model was also found, with relative SWB level accounting for a small (1% unique variance) but significant portion of the variance in Time 2 GPA. Findings also showed that the "complete mental health" group faired best overall and showed the least deterioration in GPA from Time 1 to Time 2.

Generally, Suldo and Shaffer (2008) noted several robust group distinctions in terms of outcomes that failed to replicate to the same degree when investigated a year later (Suldo et al., 2011). These findings indicate support for the possibility that more contact with positive aspects of functioning (e.g., social support, frequent positive emotions) may help attenuate the effects of PTH. Moreover, Suldo and colleagues suggest that including measures of youths' wellbeing within universal screening protocols may show incremental validity in identifying and prioritizing youth with lower or higher levels of mental health risk. Yet, given these somewhat discrepant findings, additional study is important to help clarify the nature of dual-factor mental health among youth, especially as it might apply for the purposes of mental health screening in schools.

The Current Study

Given the importance of universal mental health screening in schools and considering the evidence that a dual-factor mental health assessment framework may offer incremental validity for predicting student outcomes over and above traditional PTH assessment alone (e.g.,

Greenspoon & Saklofske, 2001; Suldo & Shaffer, 2008; Suldo et al., 2011), pursuing further research in this area may prove fruitful. Although, progress is somewhat hindered by a relative lack of appropriate and technically adequate WB screening instruments. Additionally, those that are currently available are understudied, leaving much work still to be done in understanding the role WB assessment can play in mental health screening.

Two screeners that have received some research attention include the elementary student self-report Positive Experiences at School Scale (PEASS; Furlong, You, Renshaw, O'Malley, & Rebelez, 2013) and the adolescent self-report Student Subjective Wellbeing Questionnaire (SSWQ; Renshaw, Long, & Cook, 2014). The developers of these instruments aimed to create assessment tools that were brief, measured multiple dimensions of WB, used domain-specific item wording, and were comprised of items unique to the school setting. These instruments were developed in part to address the lack of school-specific, empirically-backed youth wellbeing screeners that could tie directly in with MTSS models in school systems (Renshaw et al., 2014).

Though the PEASS (Furlong et al., 2013) and SSWQ (Renshaw et al., 2014; Renshaw, 2015) seem appropriate and have demonstrated technical adequacy as brief self-report measures that might be used for screening purposes, what has yet to be developed is an appropriate and technically adequate teacher-report instrument for screening student WB. The availability of teacher-report instruments may be desirable over self-report in situations where a self-report methodology would be a barrier to gathering useful screening data. For instance, teacher-report allows assessment of students in the educational context who may be too young and lack the self-awareness to complete self-reports with fidelity. Teacher-report also involves a more feasible data-collection procedure for elementary school settings, as it takes less time away from student learning and offers a common perspective for all student behavior within a class.

Past research has suggested that the correspondence between youth self-reports and teacher-reports of student mental health phenomena show a moderate association with each other (e.g., Earhart Jr. et al., 2009). While this does leave a substantial amount of variance in student experience left unaccounted for, teacher-reports may nonetheless be a sufficient data collection approach to identify risk as a first gate in universal mental health screening (Miller et al., 2015). Motivated by this lack of an empirically-validated school-specific teacher-report screener for student WB, the present study involved the initial development and validation efforts for such an instrument: the Student Wellbeing Teacher-Report Scales (SWTRS).

While development of a teacher-report version of the PEASS or SSWQ may seem like a logical route to crafting a teacher-report screener of student WB, the PEASS and SSWQ were specifically intended for older students and have a differential theoretical structure than what was used in the present study to develop the SWTRS. The theoretical conceptualization of student WB underlying the SWTRS drew from the "feeling good" and "functioning well" model of wellbeing, which was described above, as it has received empirical support in studies with older youth (Keyes, 2006) and broadly aligns with standard views of mental health from a psychodiagnostic perspective. For instance, in order to meet criteria for major depressive disorder in *The Diagnostic and Statistical Manual of Mental Disorders* (5^{th} ed.; American Psychiatric Association, 2013), an individual must show relatively persistent depressed mood and impairment in their typical level of adaptive daily functioning. Indeed, these symptom categories offer an example of how "feeling bad" and "functioning poorly" are key features of mental health problems, suggesting that "feeling good" and "functioning well" in life are key features of positive mental health. It is proposed that these features are thus foundational for understanding complete or optimal mental health from the dual-factor perspective.

Given the empirical and theoretical context sketched above, the following hypotheses were posited for the current study:

1. The latent factor structure of the newly generated item pool for the SWTRS would be composed of a positive internal experiences WB factor— "feeling good"—and a positive external experiences WB factor— "functioning well."
2. The WB factors derived from the SWTRS items will have adequate structural psychometric qualities for use as screening scales (e.g., robust factor loadings, acceptable internal-consistency).
3. The WB scales derived from the SWTRS factors will show small to moderate correlations with PTH measures.
4. The WB scales will predict important concurrent school outcomes (i.e., academic performance, days absent from class, and time on-task during class)
5. Using the WB scales in conjunction with PTH scales will show incremental validity evidence for predicting concurrent school outcomes (i.e., attendance, academic achievement, and time on-task) over and above using the PTH scales alone.

Method

Participants

The current study involved a sample of elementary school teachers' (N = 12) ratings of their students' (N = 184) behaviors at school. Teacher participants were recruited from a local urban charter elementary school and completed an informed consent procedure that was approved the by university's Institutional Review Board. Each grade-level at the school (K–5) had two teachers, all of whom participated in the study (Age: Median = 28 years, Range = 23–65; Years of teaching: Median = 4, Range = 1–25). Teachers were predominantly female (83.3%). In terms of highest degree earned, the sample was split evenly between bachelors and graduate degree holders. Half of the teachers identified their race/ethnicity as White/Caucasian (n = 6), with smaller proportions identifying as Black/African American (n = 3) or Multiracial (n = 1). Two teachers did not include their race/ethnicity. Student demographics as reported by teachers indicated the sample had a median age of 8 years old (Range = 5–13), were predominantly female (56%), and majority Black/African American (n = 181). Race/ethnicities of the remaining three students were identified by their teachers respectively as American Indian/Alaska Native, Native Hawaiian/Other Pacific Islander, and Multiracial. The median number of student behavior surveys submitted per teacher was 16.5 but showed wide variability across classes (Range = 7–21). This suggests that not all teachers completed surveys for all students in their classroom as the obtained sample contained 83.6% of the total school enrollment (i.e., 220 students; see "Data collection" for additional information).

Measures

Student Wellbeing Teacher-Report Scales (SWTRS). To assess positive aspects of student mental health, a new item pool of school-specific WB behaviors was drafted, refined,

and used for predictive analysis of concurrent school outcomes. See "Item pool creation process" in the "Procedures" section below for details on the scale development method for the SWTRS. Teacher respondents were asked to indicate how frequently each student displayed the school-specific WB behaviors over the previous two months. These items were followed by response options arranged along a four-point relative-frequency based scale, ranging from 0 = *Almost Never* to 3 = *Almost Always*. The original SWTRS item pool contained 54 distinct items and ultimately was reduced to two subscales containing six items each.

Student Internalizing and Externalizing Behavior Scales (SIBS and SEBS). The SIBS and SEBS are a complementary pair of teacher-report screeners intended for assessing negative aspects of student mental health. These scales are composed of seven items each, pertaining to the observable behavioral manifestations of student internalizing problems (e.g., "clings to adults", "withdrawn") and externalizing problems (e.g., "gets angry easily", "disrupts class activities"). Teachers were asked to rate how often the behaviors of interest occurred for each of their students over the previous two months. Item response options were arranged along a four-point relative-frequency based scale, ranging from 0 = *Never* to 3 = *Frequently/Almost Always.* Previous research with the SIBS and SEBS suggested that elementary school teachers could complete all 14 items for every student in a class of 25 in approximately 15–20 minutes (Cook, 2013). Such short completion times minimize the burden on the teachers and add markedly to administration feasibly (Glover & Albers, 2007). Scores derived from the SIBS and SEBS scales have strongly correlated with corresponding omnibus measures of internalizing and externalizing behaviors, moderately correlated with omnibus measures of the other construct (i.e., externalizing with internalizing), and shown adequate internal scale reliability (Cook et al., 2011; Cook, 2013). Optimal cut points were derived from receiver operating characteristic

(ROC) curve analyses and suggested that scores at or above 8 on the SIBS and at or above 9 on the SEBS marked the threshold for clinical risk (Cook, 2013). Internal consistency estimates from the present sample were above .70 for both the SIBS (α = .78) and SEBS (α = .91) scales. Descriptive summaries for SIBS and SEBS total scores are displayed in Table 1.

Table 1. Scale Descriptive Statistics for the SIBS, SEBS, AES, and SPS

Scale	Items	Min, Max	Median	Mean	SD	Q1, Q3	α	r	Skew.	Kurt.
SIBS	7	0, 17	7	7.10	4.19	4, 10	.78	.34	0.26	-0.67
SEBS	7	0, 21	9	8.76	5.98	3, 14	.91	.61	0.11	-1.21
AES	6	0, 18	11	10.74	5.08	6, 15	.93	.69	-0.07	-1.12
SPS	6	1, 18	12	11.98	4.74	8, 17	.91	.66	-0.28	-1.04

Note. Q1, Q3 = first and third quartile; *r* = average inter-item correlation; Skew. = Skewness; Kurt. = Kurtosis

Concurrent school outcomes. Multiple domains of student behavior and school functioning were measured as concurrent outcome variables in this study. These outcomes included academic achievement in (a) English language arts (ELA) and (b) math, (c) school attendance, and (d) time on-task in the classroom. These variables were selected based on their substantive worth in evaluating student success, as they are key indicators valued by teachers and administrators. Due to feasibility concerns related to linking anonymized student outcome data from school records to their teacher-rated behavior scores, the teachers provided estimates for all concurrent student outcomes. ELA and math achievement were estimated with single items that read, "In the past two months, how well has the student performed in English Language Arts/Math?", followed by a five-point response scale ranging from 1 = *Far below grade level* to 5 = *Far above grade level.* Attendance was measured with a single open-response item that read, "In the past two months, about how many full days of school has the student missed?**"** Time on-task was measured with an item that read, "In the past two months, about what percent of time was the student on-task during class?", followed by a ten-point response option scale ranging

from 1 = *0–10%* to 10 = *91–100%*. Descriptive statistics for each concurrent outcome domain are displayed in Table 2.

Table 2. Descriptive Statistics for Concurrent Outcome Variables Time On-Task, Absences, and Math and ELA Achievement

Variable	Min, Max	Median	Mean	SD	Q1, Q3	Skew.	Kurt.
Time On-Task	1, 10	8	7.50	2.39	6, 9	-0.91	2.92
Absences	0, 15	2	2.62	2.25	1, 3	1.90	8.54
Math Achievement	1, 5	3	2.69	0.99	2, 3	0.19	2.80
ELA Achievement	1, 5	3	2.62	1.02	2, 3	0.25	2.84

Note. Q1, Q3 = first and third quartile; Skew. = Skewness; Kurt. = Kurtosis

Procedure

Item pool creation process. Creation of the WB item pool followed procedures similar to those Renshaw, Long, and Cook (2014) used in developing the SSWQ and was further informed by considerations from standard texts on scale development (e.g., DeVellis, 2012; Clark & Watson, 1995) and potential treatment utility of assessment procedures (e.g., Hayes, Nelson, & Jarrett, 1987). Items were generated to reflect the hypothesized student WB dimensions "feeling good" and "functioning well" as they are represented in the literature and specifically tailored to the school environment. As an additional consideration, drafted WB items had to measure behaviors that were incompatible with those found on the SIBS and SEBS to further distinguish the PTH and WB variables. For example, the item "fights or argues with peers" on the SEBS had a complementary WB item "gets along well with classmates" that was both indicative of "functioning well" and impossible to perform simultaneously with fighting and arguing. Given that the items are intended for teacher-report, rather than student self-report, item wording targeted behaviors that could be directly observed by a teacher informant, similar to the items from the SIBS and SEBS. Two to four incompatible WB behavior items were drafted per each of the 14 total SIBS and SEBS items, resulting in the initial 54-item pool.

After item drafting, five content experts in youth wellbeing and/or school MH screening, who were all tenured professors in school psychology training programs within research-intensive universities, agreed to review the initial SWTRS item pool and rate each item in terms of (a) the construct they believed the item was most closely associated with—"feeling good," "functioning well," "both," or "neither"; (b) how sure they were of this categorization—"not very sure," "pretty sure," or "very sure"; and (c) how relevant they believed the item was to the construct they suggested it is associated with—"low relevance," "mostly relevant," or "highly relevant." The experts were given the following operational definitions to consider—"Feeling Good: Teacher's perception that a student experiences positive emotions or affective states" and "Functioning Well: Teacher's perception that student behavior is consistent with academic and social success at school." Experts were also given the option to include narrative comments about the items if appropriate. Considering this expert feedback, the item pool was then edited into a reduced and revised form. Items were removed from the pool if at least three of the five experts were "pretty sure" or "very sure" that (a) the item related to "both" or "neither" construct or (b) the item had "low relevance" to the construct they selected. The revised item pool ultimately contained 36 items.

Data collection. Teachers were asked to complete informant-report forms concerning student wellbeing behaviors (i.e., SWTRS item pool), problem behaviors (i.e., SIBS, SEBS), and the concurrent school outcomes (see "Measures") for each student in their class. Data was collected electronically using a secure online survey. Teachers were randomly assigned a letter A–L to use as both their personal anonymized identifier for their demographic information as well as the identifier for each student in their class so that student data could be appropriately clustered by teacher. Class rosters were prepared by school office staff and distributed to

teachers prior to the beginning of data collection to aid respondents in working systematically through each of their students, lessening the likelihood of a student being inadvertently excluded. No identifying student information was solicited or reported in the online survey.

Data were gathered primarily (n = 144) at a single time point at the elementary school during a one hour block normally scheduled for professional development with the author present throughout to explain the procedure and address questions if they arose. Teachers completed electronic versions of the survey using a secure online server. As an incentive for participation, all 12 teachers were entered in to a raffle at the end of the data collection period to win one of five gift cards. Some teachers did not have time to complete the survey for all of their students in the hour allotted. As a result, these teachers were allowed to complete the electronic surveys on their own during the remainder of the week. Thirty-four additional surveys were completed by midnight the following day and six more by the end of the week, completing the final data set used in the analyses.

Preliminary Analyses

All statistical analyses were conducted with IBM SPSS Statistics 23 and R statistical environment (R Core Team, 2016). As a first step, several data manipulations were performed to "tidy" the dataset prior to primary analysis in accordance with recommendations from Wickham (2014). Subsequently, preliminary analyses were conducted to explore the descriptive qualities of the data set. These preliminary analyses included inspecting visual and statistical summaries of all variables to detect aberrant data points or missing values and manually correcting any obvious data entry errors. Apart from two nonresponses to teacher race/ethnicity, no other data were missing. Further, no data points showed significant influence on the modeled data using the Mahalanobis distances procedure. During data collection, one fifth-grade teacher entered

incorrect anonymized identification codes for some of their students, making it unclear which student data were reported by which of the two fifth-grade teachers. A recoding procedure was used to identify students in the dataset who were (a) at least 11 years old and (b) did not have a code associated with the other fifth-grade teacher, who correctly entered all data. These identified cases were then reassigned their correct teacher code.

Primary Analyses

SWTRS latent structure. The first stage of the primary analyses involved an exploratory factor analysis (EFA) of the 36 SWTRS items refined from the expert review. The described procedure was based on methods used in similar scale development research (e.g., Furlong, You, Renshaw, O'Malley, & Rebelez, 2013). The purpose of the EFA was twofold: (a) to understand the latent factor structure of the SWTRS item pool and (b) to identify items for removal from the pool given that they not adhere to a latent factor in a statistically or theoretically meaningful way. Factors from the data were revealed through a factor extraction method. Because the item pool was significantly non-normal, the most appropriate extraction method was principal axis factoring (Field, 2013). A factor rotation procedure was employed to help the interpretability of the factor structure (Bryant & Yarnold, 1995). Given that latent variables in psychological research are usually correlated, an oblique factor rotation approach, direct oblimin, was most appropriate to account for this relation (Field, 2013).

Before inspecting the factor structure output, the number of preliminary indices were checked to ensure factor loading estimates were sufficiently stable and therefore interpretable. Field (2013) outlined the most important metrics and criteria to check, which include (a) ensuring that the matrix determinant is > 0, (b) the overall Kaiser-Meyer-Olkin (KMO) sampling adequacy statistic is > .50, (c) diagonal elements of the anti-image matrix are all > .50, (d) off-

diagonal estimates of the anti-image matrix are small, (e) Bartlett's Test of Sphericity is significant at $p < .05$, and (f) all communalities are > .50. Any item with a communality < .50 was a candidate for removal from the item pool. If all of the above criteria were met, interpretation of factors and item factor loadings was permissible.

EFA output includes several estimates of eigenvalues that each represent a potentially unique factor underlying the analyzed items. These eigenvalues must be interpreted in several ways to help determine the appropriate number of meaningful factors present in the data. One of the interpretive methods is a parallel analysis (Bryant & Yarnold, 1995; Whitley & Kite, 2012). The parallel analysis involves a Monte Carlo procedure that estimates the maximum eigenvalue that is likely to be obtained by chance alone for a given sample size. The number of eigenvalues larger than this estimate represent the number of potentially substantive latent factors in the model. Amount of variance extracted will be noted by summing the estimates of all eigenvalues above the result of the parallel analysis to get an indication of the amount of variance meaningfully accounted for by the EFA. A visual analysis of the scree plot—a plot of each of the extracted eigenvalues ordered from highest to lowest—is another recommended approach to help determine the number of factors (Field, 2013; Whitley & Kite, 2012). The scree plot should be interpreted by finding the point where the slope of the plotted eigenvalues changes significantly, and retain the number of eigenvalues above this break line as potential factors.

After determining the number of statistically appropriate factors from inspection of the eigenvalues, interpretation of item factor loadings in the pattern matrix output was next. An item was considered for removal from the item pool if (a) the factor loading was < .30, (b) it loaded on two or more factors > .30, or (c) it did not load onto any factor in a theoretically meaningful way (Field, 2013). Additional EFA were conducted following removal of weak functioning items

until all above conditions were satisfied. Because the items were specifically drafted to represent two hypothesized wellbeing constructs (i.e., "feeling good" and "functioning well"), it was assumed that the EFA would indicate the presence of two distinct factors from which a shortened screener-length measure could be derived (see Greco, Lambert, and Baer's [2008] use of a similar item reduction procedure in their development of the 8-item short-form of the Avoidance and Fusion Questionnaire for Youth).

SWTRS scale descriptive statistics. Several descriptive indices of scores derived from the SWTRS were evaluated. These included a count of the number of items that were ultimately retained for the new screeners, minimum and maximum scale scores, scale median, mean, and standard deviation, interquartile range, skewness and kurtosis, average inter-item correlations, and internal consistency estimates. Skewness and kurtosis estimates $\leq |3.0|$ were considered adequately normal based on the criteria specified by D'Agostino, Belanger, and D'Agostino Jr. (1990). Average inter-item correlation $r > .30$ (Field, 2013) was considered sufficiently large and internal consistency estimates (Cronbach's alpha) > .70 was the threshold for adequate reliability.

SWTRS construct validity. To investigate the construct validity of the scales derived from the SWTRS, a series of bivariate correlations were conducted between all the predictor variables (i.e., scores on the SIBS, SEBS, and SWTRS scales) and the concurrent school-related outcomes. Pearson's *r* was calculated to assess the correlations between each variable. Small to moderate negative correlations between the SWTRS scores and the SIBS and SEBS scores were predicted given that the dual-factor theory suggests PTH and WB are distinct yet related constructs (e.g., Suldo & Shaffer, 2008). Similarly, positive correlations were predicted for the relations among SWTRS scores and positive school outcomes (i.e., math and ELA achievement,

time on-task) and a negative relation with number of absences. Discrimination of the SWTRS scores from the SIBS and SEBS was also tested by entering all PTH and WB scale scores into an additional EFA with the prediction that four related but structurally unique factors would emerge.

SWTRS concurrent and incremental validity. Due to the hierarchical arrangement of the collected data, where student behavior ratings were nested within teacher respondents, multilevel modeling (MLM) procedures were utilized for all concurrent prediction analyses. Analyzing these data with MLM offered several advantages over traditional multiple regression approaches. Such advantages included the ability to calculate student-level variance separately from the variance at the class-level and for problematic patterns in the dataset (e.g., unequal sample sizes within classes, non-independence of observation) to be explicitly modeled allowing for greater estimate accuracy (Raudenbush & Bryk, 2002; Finch, Bolin, & Kelley, 2014; Huta, 2014).

The MLM approach was informed by recommendations from Hox (2010), who suggested a method in which model terms are progressively added, tested for significant model fit contribution, and subsequently retained or removed based on the result of chi-squared deviance tests. This procedure involved six modeling stages for each concurrent student outcome of interest (i.e., percent of time on-task, absences, math and reading achievement). All MLM analyses were conducted in R with the nlme package (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2016).

The first stage tested the *random intercept model,* or *null model*, which included only the outcome variable without predictors while allowing the model intercepts to vary randomly across the contextual or cluster variable. Teacher raters, or classrooms, were considered the cluster

variable for this study. This model produced an estimate of how much variability there is between average scores on the outcome variable across teachers in the population as indicated by the magnitude of the intraclass correlation (ICC).

Stage two tested the *level-1 fixed effects models* where student-level predictors were added in successive blocks. Block 1 included the fixed PTH predictors, SIBS and SEBS. Block 2 then added the new fixed WB scale predictors to test incremental validity of the dual-factor model. Consistent with the recommendation of Enders and Tofighi (2007), all level-1 predictor variables were centered within clusters to enhance interpretability of slope variance estimates and relations between level-1 predictors and the outcome variable. *Level-2 fixed effects models* followed in the next stage and involved building cluster means for the outcome variable back in to the models to restore between group variance lost by the centering procedure.

Once all desired fixed effect predictors were included, stage four tested the improvement in model fit when allowing the model slopes between each predictor variable and the outcome to vary randomly. Each predictor slope was tested individually. Once slope variability was tested for each predictor, all model slopes that contributed significantly to model fit were retained in the *full random model.* In stage five, after identifying the preferred random model, the contribution to model fit for each predictor variable was evaluated individually. In the interest of creating theoretically parsimonious models, only the predictors with significant unique contribution in explaining the outcome were retained for the final *reduced model*. Finally, stage six involved modeling the contribution of the SWTRS predictors alone to test for concurrent validity evidence.

There is no agreed upon single indicator used to determine the strength of fit for multilevel models. The commonly suggested approach involves interpretation of a variety of fit

indices to inspect relative changes in overall fit at each modeling stage and identify the best fitting explanatory model (Hox, 2010; Finch et al., 2014). A maximum likelihood estimation approach was selected for the MLM analyses. This method allowed for changes in the log likelihood fit statistic to be examined with a chi-square deviance test to detect if the magnitude of change from a simpler model to a more complex model was statistically significant. AIC and BIC fit indices were also inspected in this study. These are similar to the log likelihood statistic in that smaller values indicate better fit relative to other models. Conversely, these statistics are distinguished from the log likelihood estimates in that they inflate the estimate when model terms are included that do not make sufficiently large contributions to model fit. Of these two indices, BIC corrects the estimate more harshly than AIC.

Additionally, changes in the ICC and level-1 and level-2 pseudo R^2 were compared across models. It should be noted that the pseudo R^2 statistics used here are not the same as the more traditional R^2 estimates found in multiple regression. The R^2 values calculated for this study more accurately reflect the estimated proportion of variance in the outcome variable accounted for by a given model at level-1 and level-2, respectively. Importantly, these values should only be considered approximations of explained variance, as random slopes included in the model may bias the estimates to a small degree (Snijders & Bosker, 1999). Nonetheless, these statistics can be useful for identifying patterns across models.

The formulas used to calculate R^2 values follow the recommendations of Snijders and Bosker (1999). Calculation of level-1 R^2 values used the following formula:

$$R_1^2 = \frac{\sigma_1^2 + \tau_1^2}{\sigma_0^2 + \tau_0^2}$$

Where σ_0^2 and σ_1^2 are the level-1 error residuals for the random intercept model and the comparison model, respectively. The terms τ_0^2 and τ_1^2 indicate the intercept variance estimates

for the random intercept model and the comparison model, respectively. Calculation of level-2 R^2 values followed a similar formula:

$$R_2^2 = \frac{\sigma_1^2/B + \tau_1^2}{\sigma_0^2/B + \tau_0^2}$$

Where the model terms are the same as in the level-1 formula with the addition of *B*, which represents the average number of units per level-2 cluster. In this case, *B* was the average number of students per teacher-rater.

Results

SWTRS Latent Structure

Considering the expert feedback concerning the initial WB item pool, 18 of the 54 items were removed due to broad disagreement or poor ratings across reviewers. Testing of the normality assumption of all remaining items revealed significant multivariate non-normality. Investigation of the factor structure of the remaining 36 items proceeded with a series of principal axis factoring EFA with a direct oblimin rotation. Results of the first analysis showed strong Kaiser-Meyer-Olkin (KMO) sampling adequacy (.96) and no consequential multicollinearity (matrix determinant > 0), but some extracted item communalities were below the .50 minimum (h^2 range = .33–.80). Inspection of the factor eigenvalues and visual scree plot analysis suggested that a two-factor solution was the best fit for these items and collectively accounted for 64.23% of the total variance. Inspection of item content in relation to factor loadings suggested that the two-factor solution did not appear to align with the hypothesized "feeling good" and "functioning well" constructs but rather better represented two more context-specific constructs: a prosocial behavior factor (Student Prosociality Scale or SPS; 24 items, λ range = .47–.98) and an academic engagement factor (Academic Engagement Scale or AES; 12 items, λ range = .63–.92). Internal consistency estimates for both scales were very high (SPS α = .97; AES α = .96), suggesting possible redundancy among the items. Four items had cross-loadings above the .30 threshold.

Several additional EFA followed to clarify these results and test if removal of weaker functioning and conceptually redundant items would lead to substantive changes in the underlying factor structure. The next EFA investigated the structure of the strongest nine items from each of the two factors (18 items total) that seemed to align conceptually with the

hypothesized “feeling good” and “functioning well” (FG/FW) factors. Results indicated that despite these additional efforts to achieve a tenable FG/FW model, the items still fit more closely with a two-factor SPS and AES structure. This was also true for the subsequent 16-item EFA that excluded the two weakest items from the 18-item model. The consistency with which the SPS and AES factors appeared in each EFA suggested that the hypothesized FG/FW model was an inappropriate latent structure for these data.

Because student prosociality and academic engagement are nonetheless important indicators of WB at school, subsequent analyses used this alternative model to develop workable scales for teacher-report containing as few items as possible. The final EFA revealed a two-factor solution composed of the 12 strongest items from the 16-item model—six each from SPS and AES—and showed uniformly robust model fit indices—KMO = .93, Determinant > 0, h^2 range = .60–.73; SPS λ range = .67–.91, AES λ range = .74–.87; Cumulative variance explained = 74.10%; Factor correlation ϕ = .60, large effect. Additional EFA models containing fewer than 12 items were attempted but all yielded a single factor solution and substantially weaker structural fit, resulting in their rejection. Considering the series of EFA results together, the 12-item SPS and AES model (see Table 3) was ultimately retained as the preferred measurement model for the SWTRS, as it achieved the best balance of conceptual coherence, empirical strength, and brevity.

SWTRS Scales Descriptive Statistics

Descriptive statistics for the finalized SWTRS scales, AES and SPS, are found in Table 1.

Table 3. EFA Pattern Matrix Results for The Two-Factor SWTRS Measurement Model

	Factor Loadings (λ)	
Item	AES (ξ_1)	SPS (ξ_2)
Comfortable working independently.	0.874	-0.043
Participates meaningfully in class.	0.857	0.022
Inquisitive/interested in learning new things.	0.841	-0.046
Confident with new challenging material.	0.833	-0.033
Engaged in learning.	0.776	0.131
Shows excitement for class activities.	0.735	0.056
Friendly with classmates.	-0.068	0.926
Approachable/easy to get along with.	0.017	0.853
Self-control when frustrated.	0.037	0.818
Peaceful during class.	-0.007	0.804
Classmates respectful to them.	-0.036	0.758
Obeys class rules.	0.164	0.658
Eigenvalues	6.898	1.876
% variance	57.59	15.63

SWTRS Construct Validity

Results from the first test of construct validity, correlating the SIBS and SEBS total scores with SPS and AES total scores, showed associations in the moderate range between the SIBS and each of the SWTRS scales (SPS r = -.46, p < .01; AES r = -.47, p < .01). Correlations in the large (AES r = -.55, p < .01) and very large (SPS r = -.90, p < .01) ranges were found between the SEBS and the SWTRS scales. Next, an additional EFA was conducted including all SIBS, SEBS, SPS, and AES items together. Results revealed a three-factor model (KMO = .93, Determinant > 0, h^2 range = .10–.76; Cumulative variance explained = 59.08%) comprised of

Factor 1 (ξ_1): all seven SEBS items, all six SPS items, and the "bullied by peers" SIBS item ($|\lambda|$ range = .46–.86); Factor 2 (ξ_2): all six AES items (λ range = .72–.83); and Factor 3 (ξ_3): five SIBS items (λ range = .40–.76). The pattern matrix contained no significant cross-loadings but the SIBS item "clings to adult" was the only item with factor loadings < .30 on all three factors. Factor correlations were in the moderate (ξ_1 and ξ_3 ϕ = -.35; ξ_2 and ξ_3 ϕ = -.31) and large (ξ_1 and ξ_2 ϕ = .54) ranges.

Correlations among all PTH and WB scale total scores and measured concurrent outcome variables are found in Table 4. Both SWTRS scores showed positive correlations with time on-task and academic achievement as well as negative correlations with the number of absences as predicted; although, the SPS and absences correlation was the only non-significant relation. The SIBS and SEBS scores showed the opposite directionality of association with all outcome variables while SEBS and absences showed the only non-significant relation. The magnitude of correlation between the AES and the school-outcomes was small for absences but large for the other three variable. The AES showed the strongest associations with all outcome variables over SIBS, SEBS, or SPS. Apart from the very strong negative correlation between the SEBS and SPS, all PTH and WB correlation magnitudes were on the upper end of moderate to the low end of large. These results are largely consistent with the hypotheses and help demonstrate the validity of the dual-factor model as both PTH and WB behaviors showed meaningful associations with valued school outcomes and mid-range correlations with each other.

SWTRS Concurrent and Incremental Validity

Time On-Task. Residual errors from the multilevel models predicting percent of time on-task were visually inspected and showed adequately normal distribution and homoscedastic variance. Table 5 shows model fit indices for increasingly complex multilevel models. Table 6

Table 4. Correlation Matrix of All Predictor and Outcome Variables

Variable	1.	2.	3.	4.	5.	6.	7.	8.
1. Time On-Task	1							
2. ELA Performance	.486**	1						
3. Math Performance	.441**	.731**	1					
4. Absences	-.280**	-.240**	-.226**	1				
5. SIBS	-.375**	-.267**	-.184*	.202**	1			
6. SEBS	-.525**	-.261**	-.161*	.120	.530**	1		
7. SPS	.542**	.247**	.161*	-.078	-.461**	-.897**	1	
8. AES	.677**	.520**	.533**	-.216**	-.472**	-.547**	.589**	1

Note. Pearson correlation coefficient effect size interpretation: $r > .10$ = small, $r > .30$ = medium, $r > .50$ = large; *$p < .05$, **$p < .01$

Table 5. Fit Comparisons for Multilevel Models Predicting Percent of Time On-Task

Model Description	Model Number	Model Comparison	df	AIC	BIC	LL	LL Ratio	p
Fixed Intercept	1	--	2	845.57	852.00	-420.79	--	--
Random Intercept	2	1 v. 2	3	836.48	846.12	-415.24	11.09	< .001
Level-1 Predictors: PTH	3	2 v. 3	5	783.97	800.05	-386.99	56.50	< .001
Level-1 Predictors: PTH + WB	4	3 v. 4	7	710.49	733.00	-348.85	77.48	< .001
Level-2 Predictors: Group means	5	4 v. 5	11	708.81	744.18	-343.41	9.69	.046
Random slope: SIBS	6	5 v. 6	13	711.18	752.98	-342.59	1.63	.443
Random slope: SEBS	7	5 v. 7	13	705.95	747.75	-339.98	6.86	.032
Random slope: AES	8	5 v. 8	13	700.23	742.02	-337.11	12.58	.002
Random slope: SPS	9	5 v. 9	13	703.73	745.53	-338.87	9.08	.011
Full Random Model	10	5 v. 10	20	712.66	776.96	-336.33	14.16	.117
Adjusted Full Random Model*	11	5 v. 11	16	704.90	756.34	-336.45	13.91	.016
Reduced Model	12	2 v. 12	10	703.18	735.33	-341.59	147.30	< .001

Note. LL = Log Likelihood. *The random slope for SEBS was eliminated as it made the weakest contribution to overall model fit.

Table 6. Coefficient and Effect Size Estimates for Multilevel Models Predicting Percent of Time On-Task

Model	Random Intercept Only		Level-1 Predictors: PTH		Level-1 Predictors: PTH + WB		Level-2 Predictors: Group Means		Adjusted Full Random Model		Reduced Model	
	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*
Fixed												
Intercept	7.42(.30)	.88[c]	7.42(.31)	.88[c]	7.41(.31)	.88[c]	6.08(7.87)	.06	7.61(6.97)	.08	7.41(.31)	.88[c]
SIBS			-0.04(.04)	-.07	0.06(.04)	.13	0.06(.04)	.12	0.06(.03)	.12		
SEBS			-0.19(.03)	-.45[c]	-0.03(.05)	-.05	-0.03(.05)	-.05	-0.01(.05)	-.01		
AES					0.27(.03)	.56[c]	0.27(.03)	.56[c]	0.28(.04)	.50[c]	0.27(.03)	.50[c]
SPS					0.08(.06)	.10	0.08(.06)	.10	0.12(.06)	.15[a]	0.11(.03)	.25[c]
SIBS means							-0.47(.19)	-.68[a]	-0.38(.16)	-.66		
SEBS means							0.15(.32)	.17	0.11(.28)	.15		
AES means							0.30(.16)	.57	0.22(.15)	.50		
SPS means							0.02(.45)	.02	-0.07(.39)	-.07		
Random												
Level-1 residual variance	4.75		3.56		2.27		2.27		2.03		2.06	
Level-2 intercept variance	0.75		0.85		0.95		0.32		0.40		0.95	
Slope variance: AES	--		--		--		--		0.005		0.006	
Slope variance: SPS	--		--		--		--		0.002		0.001	
Model Effect Size												
Level-1 Pseudo R^2	--		.23		.43		.54		.57		.47	
Level-2 Pseudo R^2	--		.28		.55		.53		.58		.59	
ICC	.13		.19		.29		.12		.16		.32	

Note. ICC = intraclass correlation. The *t* statistic and associated degrees of freedom for each fixed effect predictor was converted to correlation effect size estimate *r* to enhance interpretability; $r > .10$ = small, $r > .30$ = medium, $r > .50$ = large; [a] $p < .05$, [b] $p < .01$, [c] $p < .001$.

shows the comparative fixed effect predictor estimates, random variance components, and omnibus model effect size estimates at level-1 (students) and level-2 (teachers) for each successive model. The first stage of modeling tested the random intercept model and revealed significant variability in average time on-task across teacher raters and was thus retained. Specifically, the ICC suggested that 13% of the variability of time on-task estimates were accounted for purely by the teacher contextual variable.

The next stage tested the incremental model improvement as each block of level-1 and -2 fixed effect predictors were entered beginning with inclusion of the SIBS and SEBS total scores. Adding these level-1 PTH variables resulted in significant overall model improvement, but SEBS (r = -.45, p < .001; moderate effect) was the only significant individual predictor. AES and SPS total scores were then added in the next block as additional level-1 fixed effect predictors. Adding this block resulted in significant improvement in model fit over and above SIBS and SEBS alone. Furthermore, including the SWTRS variables resulted in AES (r = .56, p < .001; large effect) now being the only significant predictor of time on-task. Level-2 teacher means for each of the four level-1 predictors were then added as the last fixed effect block. This also resulted in significant overall model improvement with both AES (r = .56, p < .001; large effect) and SIBS means (r = -.68, p < .05; large effect) as significant individual predictors. This model containing all level-1 and -2 fixed effects was deemed the preferred fixed effects model.

Random slopes for the relation between each level-1 predictor and time on-task were tested individually for significant variability across teachers in the next modeling stage. SEBS (slope variance = 0.004), AES (slope variance = 0.006), and SPS (slope variance = 0.003) each showed significant variability across teachers but when random slopes for all three variables were included as a collective block, there was not significant improvement in overall model fit.

Examination of each variable's respective model fit statistics revealed the random slope term for SEBS contributed the least to overall fit improvement and was thus removed from the model. Retaining only the AES (slope variance = 0.005) and SPS (slope variance = 0.002) random slope terms resulted in significant overall model improvement and was thus considered the preferred (adjusted) full random model. The AES (r = .50, p < .001; large effect) and SPS (r = .15, p < .05; small effect) fixed effect variables were the only significant individual predictors of time on-task in this model.

The final modeling stage involved removal of all non-significant fixed effect predictors from the adjusted full random model to test the comparative fit of a more parsimonious, reduced model. This model retained the significant AES (r = .50, p < .001; large effect) and SPS (r = .25, p < .001; small effect) level-1 predictors, and their respective random slope variance terms (AES = 0.006; SPS = 0.001). The model also resulted in significantly improved overall model fit compared to the random intercept model. Comparison of the fit indices between this model and the adjusted full random model suggested that the reduced model was preferred for predicting time on-task as it showed the strongest balance between model fit and conceptual parsimony. This reduced model also functioned as the SWTRS only model and ultimately accounted for approximately 47% of variance at level-1 and 59% at level-2.

Absences. Residual errors from the multilevel models predicting number of absences were visually inspected and showed adequately normal distribution and homoscedastic variance. Table 7 shows the model fit indices for increasingly complex multilevel models. Table 8 shows the comparative fixed effect predictor estimates, random variance components, and omnibus model effect size estimates at level-1 (students) and level-2 (teachers) for each successive model. The first stage of modeling revealed that 32% of the variability in number of absences was

Table 7. Fit Comparisons for Multilevel Models Predicting Number of Absences

Model Description	Model Number	Model Comparison	df	AIC	BIC	LL	LL Ratio	p
Fixed Intercept	1	--	2	823.38	829.81	-409.69	--	--
Random Intercept	2	1 v. 2	3	783.45	793.09	-388.73	41.93	< .001
Level-1 Predictors: PTH	3	2 v. 3	5	780.98	797.05	-385.49	6.47	.039
Level-1 Predictors: PTH + WB	4	3 v. 4	7	780.76	803.27	-383.38	4.21	.122
Level-2 Predictors: Group means	5	4 v. 5	11	779.29	814.65	-378.64	9.47	.050
Random slope: SIBS	6	5 v. 6	13	780.10	821.89	-377.05	3.19	.203
Random slope: SEBS	7	5 v. 7	13	793.29	825.09	-378.65	0.00	.999
Random slope: AES	8	5 v. 8	13	783.23	825.02	-378.62	0.06	.971
Random slope: SPS	9	5 v. 9	13	793.06	824.86	-379.53	0.23	.892
Reduced Model	10	2 v. 10	4	776.14	789.00	-384.07	9.31	.002
SWTRS Only Model	11	2 v. 11	5	778.14	794.22	-384.07	9.31	.010

Note. LL = Log Likelihood.

Table 8. Coefficient and Effect Size Estimates for Multilevel Models Predicting Number of Absences

Model	Random Intercept Only		Level-1 Predictors: PTH		Level-1 Predictors: PTH + WB		Level-2 Predictors: Group Means		Reduced Model		SWTRS Only Model	
	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*
Fixed												
Intercept	2.73(.42)	.44[c]	2.73(.43)	.44[c]	2.73(.43)	.44[c]	-16.81(10.96)	-.12	2.73(.43)	.44[c]	2.73(.43)	.44[c]
SIBS			0.06(.04)	.12	0.04(.04)	.07	0.04(.04)	.07				
SEBS			0.03(.03)	.08	0.03(.06)	.04	0.03(.06)	.04				
AES					-0.07(.04)	-.16[a]	-0.07(.04)	-.15[a]	-0.08(.03)	-.23[b]	-0.08(.03)	-.19[a]
SPS					0.04(.07)	.05	0.04(.07)	.05			0.00(.04)	.00
SIBS means							0.55(.27)	.61				
SEBS means							0.29(.44)	.24				
AES means							-0.30(.23)	-.44				
SPS means							1.35(.62)	.63				
Random												
Level-1 residual variance	3.47		3.34		3.26		3.26		3.29		3.29	
Level-2 intercept variance	1.90		1.91		1.92		0.75		1.92		1.92	
Model Effect Size												
Level-1 Pseudo R^2	--		.02		.04		.25		.03		.03	
Level-2 Pseudo R^2	--		.04		.06		-.01		.05		.05	
ICC	.35		.36		.37		.19		.37		.37	

Note. ICC = intraclass correlation. The *t* statistic and associated degrees of freedom for each fixed effect predictor was converted to correlation effect size estimate *r* to enhance interpretability; $r > .10$ = small, $r > .30$ = medium, $r > .50$ = large; [a] $p < .05$, [b] $p < .01$, [c] $p < .001$.

accounted for by the teacher-rater contextual variable. The significant random intercept term was thus retained for subsequent models.

The first block of the fixed effect modeling stage showed significant model fit improvement when the level-1 SIBS and SEBS total scores were included, although neither predictor accounted for a significant amount of unique variance individually. AES and SPS total scores were then added in the next block as additional level-1 fixed effect predictors but did not contribute to significant improvement in model fit over and above SIBS and SEBS alone. Despite this, AES (r = -.16, p < .05; small effect) was the only significant individual predictor included in this model. Although inclusion of the WB block resulted in non-significant fit improvement, the variables were retained in subsequent models due to the possibility that adding level-2 group means or random slope terms could change conclusions about the predictors' explanatory contributions to the dependent variable. Level-2 teacher means for each of the for level-1 predictors were next added as the last fixed effect block. This resulted in marginally significant model improvement over the WB block with AES (r = -.15, p < .05; small effect) again being the only significant individual predictor. This model containing all level-1 and -2 fixed effects was the preferred fixed effects model.

Random slopes for the relation between each level-1 predictor and number of absences were tested individually for variability across teachers in the following stage but none resulted in significant model improvement. As a result, no preferred random model was identified for predicting number of absences.

Selecting the significant model terms from the preferred fixed effect model, a reduced model was then tested that included AES (r = -.23, p < .01; small effect) as the only predictor. Fit indices suggested that this reduced model showed significant overall model improvement

compared to the random intercept model and best balanced parsimony with statistical fit of all previously tested models. However, the reduced model accounted for much less level-1 variance (~5%) in predicting absences than did the preferred fixed effect model (~25%).

Finally, the SPS predictor was added back in to the reduced model in order to test the concurrent prediction power of the SWTRS variables alone. Although the unique contribution of SPS was extremely negligible, resulting in a somewhat less parsimonious model than the reduced model, SWTRS only nonetheless showed significant improvement in fit over the random intercept model as predicted. The amount of variance accounted for at level-1 and -2 was virtually identical to the reduced model with AES as the sole significant predictor (r = -.19, p < .05; small effect).

Math Achievement. Residual errors from the multilevel models predicting student math achievement relative to grade-level norms were visually inspected and showed adequately normal distribution and homoscedastic variance. Table 9 shows the model fit indices for increasingly complex multilevel models. Table 10 shows the comparative fixed effect predictor estimates, random variance components, and omnibus model effect size estimates at level-1 (students) and level-2 (teachers) for each successive model. The first stage of modeling indicated variability in math achievement variance across teachers was essentially 0, suggesting that subsequent models should maintain a fixed intercept yet still test for significant variability in model slopes. Note that all estimates for level-2 intercept variance and ICC in Table 10 are 0 and level-2 pseudo R^2 values are identical to level-1 values across all models. This was an artifact of keeping the model intercepts fixed. Despite this, these estimates remain in Table 10 to maintain consistency with the other tables.

Table 9. Fit Comparisons for Multilevel Models Predicting Math Performance

Model Description	Model Number	Model Comparison	df	AIC	BIC	LL	LL Ratio	p
Fixed Intercept	1	--	2	521.84	528.27	-258.92	--	--
Random Intercept	2	1 v. 2	3	523.84	533.48	-258.92	0.00	> .999
Level-1 Predictors: PTH	3	1 v. 3	4	515.11	527.97	-253.55	10.73	.005
Level-1 Predictors: PTH + WB	4	3 v. 4	6	459.02	478.31	-223.51	60.09	< .001
Level-2 Predictors: Group means	5	4 v. 5	10	461.93	494.08	-220.97	5.08	.279
Random slope: SIBS	6	4 v. 6	7	460.90	483.41	-223.45	0.12	.734
Random slope: SEBS	7	4 v. 7	7	461.02	483.52	-223.51	0.00	.967
Random slope: AES*	8	4 v. 8	7	451.14	473.65	-218.57	9.88	.002
Random slope: SPS	9	4 v. 9	7	461.02	483.52	-223.51	0.00	.966
Reduced Model	10	1 v. 10	4	451.85	464.71	-221.92	73.98	< .001
SWTRS Only Model	11	1 v. 11	5	449.01	465.08	-219.50	78.83	< .001

Note. LL = Log Likelihood. *Also the Full Random Model as AES had the only significant random slope.

Table 10. Coefficient and Effect Size Estimates for Multilevel Models Predicting Math Performance

Model	Random Intercept Only		Level-1 Predictors: PTH		Level-1 Predictors: PTH + WB		Level-2 Predictors: Group Means		Full Random Model		Reduced Model		SWTRS Only Model	
	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*
Fixed														
Intercept	2.68(.07)	.94[c]	2.68(.07)	.94[c]	2.68(.06)	.96[c]	1.07(2.69)	.03	2.68(.06)	.96[c]	2.68(.06)	.96[c]	2.68(.06)	.96[c]
SIBS			-0.04(.02)	-.14	0.01(.02)	.10	0.01(.02)	.04	0.03(.02)	.10				
SEBS			-0.02(.01)	-.11	-0.02(.03)	-.02	-0.02(.03)	-.07	-0.01(.02)	-.02				
AES					0.14(.02)	.39[c]	0.14(.02)	.53[c]	0.15(.03)	.39[c]	0.12(.02)	.35[c]	0.14(.02)	.39[c]
SPS					-0.07(.03)	-.09	-0.07(.03)	-.16[a]	-0.04(.03)	-.09			-0.04(.02)	-.17[a]
SIBS means							-0.04(.06)	-.26						
SEBS means							0.09(.11)	.30						
AES means							0.10(.05)	.57						
SPS means							0.01(.15)	.02						
Random														
Level-1 residual variance	0.98		0.92		0.66		0.65		0.58		0.60		0.59	
Level-2 intercept variance	0.00		0.00		0.00		0.00		0.00		0.00		0.00	
Slope variance: AES	--		--		--		--		0.005		0.005		0.004	
Model Effect Size														
Level-1 Pseudo R^2	--		.06		.32		.34		.41		.39		.40	
Level-2 Pseudo R^2	--		.06		.32		.34		.41		.39		.40	
ICC	.00		.00		.00		.00		.00		.00		.00	

Note. ICC = intraclass correlation. The *t* statistic and associated degrees of freedom for each fixed effect predictor was converted to correlation effect size estimate *r* to enhance interpretability; *r* > .10 = small, *r* > .30 = medium, *r* > .50 = large; [a] *p* < .05, [b] *p* < .01, [c] *p* < .001

The first block of the fixed effect modeling stage showed significant model fit improvement when the level-1 SIBS and SEBS total scores were included, although neither predictor accounted for a significant amount of unique variance individually. AES and SPS total scores were then added in the next block as additional level-1 fixed effect predictors, resulting in significant model fit improvement over and above SIBS and SEBS alone. AES (r = .53, p < .001; large effect) and SPS (r = -.16, p < .05; small effect) were the only significant individual predictors in this model. Level-2 teacher means for each of the four level-1 predictors were next added as the last fixed effect block. This inclusion of level-2 predictors did not significantly improve model fit and level-1 predictor estimates remained effectively unchanged. The model including all level-1 PTH and WB predictors, but not level-2 means, was accepted as the preferred fixed effect model.

Random slopes for the relation between each level-1 predictor and math achievement were tested individually for variability across teachers in the next modeling stage. Only the AES slopes were identified as varying significantly (slope variance = 0.005) while meaningfully improving model fit. Including this random slope term resulted in a significant, but somewhat weaker, AES fixed effect estimate (r = .39, p < .001; moderate effect) and reduced the SPS fixed effect relation with math achievement to non-significance. The variance accounted for in the model did however increase from 34% to 41%. This was the preferred random model.

The significant AES variable maintained statistical significance when entered in to the reduced model as the sole predictor (r = .35, p < .001; moderate effect). Fit indices suggested that this reduced model showed significant overall model improvement compared to the fixed intercept-only model.

Finally, reintroducing SPS into the reduced model resulted in significant improvement in model fit over the fixed intercept only model. This SWTRS only model also showed the best balance of parsimony with statistical fit of all models tested and accounted for approximately 40% of the variance. Both AES (r = .39, p < .001; moderate effect) and SPS (r = -.17, p < .05; small effect) made significant individual contributions in explaining math performance.

ELA Achievement. Residual errors from the multilevel models predicting student ELA achievement relative to grade-level norms were visually inspected and showed adequately normal distribution and homoscedastic variance. Table 11 shows the model fit indices for increasingly complex multilevel models. Table 12 shows the comparative fixed effect predictor estimates, random variance components, and omnibus model effect size estimates at level-1 (students) and level-2 (teachers) for each successive model. The first stage of modeling indicated that, like math achievement, ELA achievement variance across teachers was effectively 0, suggesting that subsequent models should maintain a fixed intercept but still test for the presence of significant slope variability. Note that all estimates for level-2 intercept variance and ICC in Table 12 are 0 and level-2 pseudo R^2 values are identical to level-1 values across all models. This was an artifact of maintaining fixed intercepts in all of the models. Despite this, these estimates were left in Table 12 to maintain consistency with the other tables.

The first block of the fixed effect modeling stage showed significant model fit improvement when the level-1 SIBS and SEBS total scores were included, with SEBS significantly predicting ELA achievement uniquely (r = -.16, p < .05; small effect). AES and SPS total scores were then added in the next block as additional level-1 fixed effect predictors resulting in significant overall model fit improvement beyond the SIBS and SEBS block alone. AES (r = .44, p < .001; moderate effect) was the only significant individual predictor in this

Table 11. Fit Comparisons for Multilevel Models Predicting ELA Performance

Model Description	Model Number	Model Comparison	df	AIC	BIC	LL	LL Ratio	p
Fixed Intercept	1	--	2	531.46	537.89	-263.73	--	--
Random Intercept	2	1 v. 2	3	533.46	543.11	-263.73	0.00	> .999
Level-1 Predictors: PTH	3	1 v. 3	4	519.39	532.25	-255.69	16.07	< .001
Level-1 Predictors: PTH + WB	4	3 v. 4	6	483.85	503.14	-235.93	39.54	< .001
Level-2 Predictors: Group means	5	4 v. 5	10	484.06	516.21	-232.03	7.79	.100
Random slope: SIBS	6	4 v. 6	7	485.02	507.53	-235.51	0.83	.363
Random slope: SEBS	7	4 v. 7	7	485.05	507.55	-235.52	0.80	.370
Random slope: AES*	8	4 v. 8	7	475.61	498.11	-230.81	10.24	.001
Random slope: SPS	9	4 v. 9	7	484.61	507.12	-235.31	1.24	.265
Reduced Model	10	1 v. 10	4	470.55	483.41	-231.27	64.91	< .001
SWTRS Only Model	11	1 v. 11	5	472.04	488.11	-231.02	65.43	< .001

Note. LL = Log Likelihood. *Also the Full Random Model as AES had the only significant random slope.

Table 12. Coefficient and Effect Size Estimates for Multilevel Models Predicting ELA Performance

Model	Random Intercept Only		Level-1 Predictors: PTH		Level-1 Predictors: PTH + WB		Level-2 Predictors: Group Means		Full Random Model		Reduced Model		SWTRS Only Model	
	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*	Estimate (SE)	*r*
Fixed														
Intercept	2.62(.07)	.94[c]	2.62(.07)	.94[c]	2.62(.07)	.95[c]	4.60(2.86)	.16	2.62(.06)	.96[c]	2.62(.06)	.96[c]	2.62(.06)	.96[c]
SIBS			-0.04(.02)	-.14	-0.00(.02)	-.00	-0.00(.02)	-.00	0.01(.02)	.05				
SEBS			-0.03(.01)	-.16[a]	-0.02(.03)	-.06	-0.02(.03)	-.06	-0.00(.03)	-.01				
AES					0.11(.02)	.44[c]	0.11(.02)	.45[c]	0.12(.03)	.34[c]	0.11(.02)	.34[c]	0.11(.02)	.34[c]
SPS					-0.04(.03)	-.09	-0.04(.03)	-.09	-0.01(.03)	-.03			-0.01(.02)	-.05
SIBS means							-0.07(.07)	-.39						
SEBS means							-0.05(.11)	-.17						
AES means							0.09(.05)	.54						
SPS means							-0.17(.16)	-.37						
Random														
Level-1 residual variance	1.03		0.94		0.76		0.73		0.67		0.68		0.68	
Level-2 intercept variance	0.00		0.00		0.00		0.00		0.00		0.00		0.00	
Slope variance: AES	--		--		--		--		0.004		0.004		0.003	
Model Effect Size														
Level-1 Pseudo R^2	--		.08		.26		.29		.34		.34		.34	
Level-2 Pseudo R^2	--		.08		.26		.29		.34		.34		.34	
ICC	.00		.00		.00		.00		.00		.00		.00	

Note. ICC = intraclass correlation. The *t* statistic and associated degrees of freedom for each fixed effect predictor was converted to correlation effect size estimate *r* to enhance interpretability; $r > .10$ = small, $r > .30$ = medium, $r > .50$ = large; [a] $p < .05$, [b] $p < .01$, [c] $p < .001$.

model. Level-2 teacher means for each of the four level-1 predictors were next added as the last fixed effect block but did not significantly improve overall model fit. Level-1 predictor estimates were again largely unchanged. The model including all level-1 PTH and WB predictors, but not level-2 means, was accepted as the preferred fixed effects model.

Random slopes for the relation between each level-1 predictor and ELA achievement were tested individually for variability across teachers in the next modeling stage but only the AES slopes were indicated as varying significantly (slope variance = 0.004) and meaningfully improving overall model fit. Including this random slope term resulted in a significant, but somewhat weaker, AES fixed effect (r = .34, p < .001; moderate effect). The variance accounted for in the model did however increase from about 29% to 34%. This was the preferred full random model.

The significant AES variable maintained statistical significance when entered in to the reduced model as the sole predictor (r = .34, p < .001; moderate effect). Fit indices suggested that this reduced model showed significant overall model improvement compared to the fixed intercept model and best balanced parsimony with statistical fit of all tested models. This reduced model accounted for approximately 34% of the variance in ELA achievement.

Reintroducing the SPS predictor into the model resulted in significant improvement overall compared to the fixed intercept model, as predicted. This SWTRS only model had the second strongest fit after the reduced model and similarly explained about 34% of the total variance. AES was again the only significant individual predictor (r = .34, p < .001; moderate effect).

Discussion

The idea that positive aspects of youth mental health should be included in school-based mental health screening is continuing to gain traction among scholars and school practitioners alike. As several studies have suggested (e.g., Renshaw & Cohen, 2014; Suldo & Shaffer, 2008), the standard unidimensional model of mental health that equates the absence of problems with the presence of wellbeing may in fact be overly simple. The dual-factor mental health model posits that mental health could be alternatively conceptualized as existing along two related but distinct continua—PTH and WB. The model offers increased nuance to our understanding of what makes for complete mental health and may also contribute to greater precision in identifying and prioritizing youth with higher or lower levels of mental health risk (Dowdy et al., 2014). For instance, while students A and B may both be experiencing significant symptoms of distress in their lives, if student A also experiences a greater frequency of wellbeing behaviors (e.g., positive peer relationships, feelings of excitement) than student B, the dual-factor model would suggest that student A may actually have better school-related outcomes due to the protective nature of the higher wellbeing.

Considering the above logic, the present study had two broad aims related to understanding youth WB in the school context. The first aim was to investigate the structural validity of teacher-report items developed to assess a theoretical model of student WB composed of "feeling good" and "functioning well" dimensions, which has historically been validated only through self-report measures (e.g., Keyes & Annas, 2009). This was examined through the initial development of an item pool of school-specific student WB behaviors and subsequent reduction of the pool to brief scales—the SWTRS—potentially suitable for use in universal mental health screening. The second aim concerned investigating the concurrent validity of SWTRS scores via correlating with their counterpart PTH scores as well as in predicting important school-related concurrent outcomes, both

with and without PTH scale scores included as additional predictor variables. Taken together, the aims of the present study were intended to provide initial evidence in support of the technical adequacy of a new brief teacher-report measure that might be useful for assessing student wellbeing within a dual-factor mental health screening framework in schools.

Structural Validity

Evidence from the series of exploratory factor analyses showed unambiguously that the FG/FW model of WB did not generalize as hypothesized to the SWTRS item pool. Nevertheless, two distinct latent factors consistently emerged in each EFA: (a) a student prosocial behavior factor (SPS; example items: "friendly with classmates," "shows self-control when frustrated") and (b) an academic engagement factor (AES; example items: "comfortable working independently," "inquisitive/interested in learning new things"). Despite not representing the FG/FW model well, the AES and SPS still appear to measure important indicators of positive mental health (Cowen & Kilmer, 2002).

In fact, a developing base of empirical work validating the Social, Academic, and Emotional Behavior Risk Screener (SAEBRS; see Kilgus, Chafouleas, & Riley-Tillman, 2013; von der Embse, Pendergast, Kilgus, & Eklund, 2015; Kilgus, Eklund, von der Embse, Taylor, & Sims, 2016) has yielded similar teacher-report scales to the SWTRS, albeit by using an alternate theoretical lens. The primary difference is that the SAEBRS development was informed in part by the theory of "academic enablers" (Kilgus et al., 2013), which suggests that display of adaptive social and academic work-related behaviors, along with withholding maladaptive behaviors, is associated with academic achievement (DiPerna, 2006; Volpe et al., 2006). In comparison, the SWTRS development was informed by the youth wellbeing literature which suggests that the measurement of positive behaviors has larger utility in understanding youth mental health beyond only predicting academic risk. While

there is overlap in the types of behaviors measured in the SAEBRS and SWTRS, the subscales are conceptually distinct enough to warrant use in differing universal screening applications depending on the school's goals. Because both measures have been associated with school-related outcomes, future research could investigate the comparative predictive validity of scores derived from both measures to empirically demonstrate if one is more strongly associated with school outcomes than another.

Considering the item content of the SWTRS in light of past research, it appears that both AES and SPS assess broad meta-constructs that each contain items related to more specific sub-domains that have been targeted in other measure development research driven by different conceptual models. For example, although there is wide disagreement among scholars concerning the definition and important features of school engagement (see Appleton, Christenson, & Furlong [2008] and Fredricks et al. [2011] for overviews of competing theories and measurement approaches), a three-factor model has received the most substantial support in the literature (Jimerson, Campos, & Greif, 2003). This tripartite model suggests that measuring (a) behavioral, (b) cognitive, and (c) emotional or affective dimensions of engagement are critical components to include when measuring the construct (Fredricks, Blumenfeld, & Harris, 2004). The use of this engagement model has precedence in past dual-factor mental health work, such as in Lyons, Huebner, and Hills' (2012) study that used measures of each of the three dimensions as outcomes predicted by subjective WB scores. Consistent with this model, the AES items appear to relate to all three of these dimensions of school engagement.

Similarly, the SPS contains item content similar to a five-factor model of social skills consisting of (a) peer relations, (b) self-management, (c) academic behaviors, (d) compliance, and (e) assertion. Support for this model came originally through meta-analysis of item content within

several social skills rating scales (Caldarella & Merrell, 1997). The factor analytic evidence in Caldarella and Merrell's study suggested that these five dimensions were the most commonly assessed domains of social skills, although considerable conceptual overlap was also suggested among the domains. Further evidence was found for this model through later qualitative research investigating which youth social skills were considered most important to children, parents, and teachers (Warnes, Sheridan, Geske, & Warnes, 2005).

In contrast to the proposed cross-context FG/FW model, where students' "feeling good" and "functioning well" behaviors are assumed to be consistent across situations, the results from the present study suggest that student WB behaviors at school may be better understood as a function of context (i.e., in social situations and during academic activities) when the teacher is the informant. As previously stated, the FG/FW model was originally developed from self-report evidence, which allows for greater reliability and validity of private behavior measurement as compared to informant-report (Merrell, 2008). Because of this, the developed SWTRS items were specifically worded in terms of observable public behaviors to lower the level of inference for teacher ratings.

Although the SWTRS latent factors did not appear as hypothesized, AES and SPS were still compared to the SIBS and SEBS to see if the scale constructs could be adequately discriminated. As before, when operating from the FG/FW framework, evidence from bivariate correlations of the PTH and WB variables that were non-trivial but not so large to conclude they were measuring the same thing was considered support for their distinction. Results from the MLM analyses that showed the addition of WB variables improved model fit in explaining the outcome over and above PTH was considered additional evidence for the dual-factor framework.

Three of the four correlations between the PTH and SWTRS scales' scores showed associations bordering between medium and large effects. These estimates were on the upper end of

what was expected from the dual-factor perspective, yet were still largely consistent with the hypothesized magnitude of relation. Moreover, these results were in line with what Earhart et al. (2009) found in their dual-factor MH study of the correlations among various student WB (i.e., hope, life satisfaction, school connectedness) and PTH indicators (i.e., student- and teacher-report BESS scores). In contrast, SPS and SEBS showed a very large negative correlation similar to the correlation found between the SABRS Social Behavior scale and the Social Skills Improvement System-Social Skills scale (Kilgus et al., 2013). The subsequent EFA that included all SIBS, SEBS, AES, and SPS items resulted in a three factor solution where the SEBS and SPS loaded together on a single factor while the AES and SIBS were mostly distinct entities. This is consistent with empirical evidence from SAEBRS structural validation research that showed the presence of three unique factors—Academic Behavior (cf. AES), Social Behavior (cf. SPS, SEBS), and Emotional Behavior (cf. SIBS) where Academic and Social Behavior scales contained both positive and negatively worded items (von der Embse et al., 2015). The unidimensionality of social skills versus social deficits is further backed by theoretical work suggesting a conceptually inverse relation between the two (cf. Caldarella & Merrell, 1997; Quay, 1986). Altogether, it was concluded that the SPS is best described as measuring a positive inverse of externalizing behavior problems rather than a unique aspect of WB like the AES.

Interestingly, the SIBS—the PTH complement to the "feeling good" factor originally hypothesized—showed at least adequate internal consistency both in past research (e.g., Cook et al., 2011) and the current sample despite also focusing exclusively on the public manifestations of internal experiences for teacher-report. It is important to note that the SIBS and SEBS items were not developed through factor analytic means, as in the present study, but only through literature review and selection of the 14 most relevant "internalizing" and "externalizing" youth problem behaviors as

rated by content experts (Cook, 2013). It is not completely clear whether substantive differences exist between the theoretical structures of youth PTH and WB or if the discrepancies in scale development methodology better account for the lack of FG/FW evidence in this study, although evidence from the SAEBRS development suggests the former (e.g., von der Embse, 2015).

Concurrent and Incremental Validity

In modeling the concurrent relations among the PTH and WB predictors and the four school outcomes—percent of time on-task during class, number of absences, math and ELA achievement—multiple patterns emerged in the results. As hypothesized, models that included only the two WB variables as predictors resulted in significant improvement in model fit over the null model for all four outcome domains, consistent with similar past research (e.g., Kim, Furlong, Dowdy, & Felix, 2014). Furthermore, these SWTRS only models were the strongest fitting overall for predicting time on-task and math achievement. Both WB indicators were significant individual predictors for the math achievement (cf. Suldo et al., 2011) and time on-task outcomes as well. SWTRS only models were also the second-best fitting for concurrently predicting absences and ELA achievement, after the reduced models, which both retained AES as the only significant individual predictor. These results broadly align with past findings that suggest significant associations of WB variables with important school-related outcomes (e.g., Lyons et al., 2012; Suldo et al., 2011).

Inclusion of the PTH block of predictors alone consistently resulted in significant improvements in model fit over the null models of all four outcome domains (cf. Kim et al., 2014). Although the PTH blocks showed significant fit for each outcome, SIBS never accounted for more than a negligible or small but non-significant portion of unique variance. On the other hand, SEBS did show a small individual effect predicting ELA achievement and a moderate effect predicting time on-task. Subsequently adding the two WB predictors to the models in a second block resulted in

significant improvements in model fit over the PTH predictors alone for all outcomes excluding number of absences. Multiple past studies corroborated this additive contribution of WB variables over and above PTH (e.g., Lyons et al., 2012; Kim et al., 2014). Despite the WB block not contributing significant improvement over the PTH block here, AES was the only significant individual predictor included in the model, accounting for a small portion of unique variance in absences. Moreover, the SWTRS only model showed superior fit over the null model as compared to the PTH block alone. This finding suggests that the contribution of WB behaviors in predicting school absences considerably diminishes when also factoring in the student's relative level of internalizing and externalizing PTH behaviors. Including the SWTRS also resulted in the significant individual contributions of SEBS scores in predicting time on-task and ELA achievement to shrink to negligible effects (cf. Lyons et al., 2012). Considering this incremental validity evidence, support for a dual-factor model was largely found for predicting achievement and engagement outcome variables beyond using PTH variables in exclusivity.

Given the evidence that the SEBS and SPS were negatively correlated but not structurally unique, it is reasonable that after SPS was included in the models, the predictor absorbed much of the unique variance previously accounted for by the SEBS alone. Despite this, these variables did not behave consistently as inverse yet equitable predictors as might be expected (e.g., Kilgus et al, 2013). First, SPS functioned as the stronger variable in predicting all four school outcomes in models that included all PTH and WB variables. Although this pattern was consistent, the advantage of SPS over SEBS was marginal for each outcome, apart from math achievement where the individual effect of SEBS was negligible while the effect of SPS was small yet significant.

Second, the directionality of the predictor estimates for both variables was actually the same for models predicting absences (positive association) and both math and ELA achievement (negative

association), yet opposite for predicting time on-task (SPS: positive association; SEBS: negative association). While it is tempting to interpret this result as suggesting that a greater frequency of both disruptive and prosocial behavior may indicate greater risk in academic achievement and school attendance in a teacher-report mental health screening context, this pattern was not found when inspecting the bivariate correlations (see Table 4) among the predictor and outcome variables. In fact, the relations between the scales and the outcomes consistently showed opposite directionality for WB versus PTH predictors (e.g., AES and SPS positively correlated with math achievement while SIBS and SEBS showed a negative correlation). The need for additional study on this phenomenon is evident in order to clarify the discrepancy in the valence of the variable relations.

Limitations and Future Directions

Although the results of the present study are interesting, these should be considered in light of a few important limitations and suggestions for future research. For instance, values of all concurrent student outcome variables were estimated solely from teacher-reports. Although convenient from a data collection standpoint, gathering all data exclusively from teacher-reports may have biased the results of the concurrent validity analyses due to the influence of common method variance (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). One approach future researchers could use to help control for this bias is to explicitly link each student's teacher-estimated PTH and WB behavior frequency to outcomes derived from other measurement sources, such as standardized test scores for achievement, school records for number of absences, or direct behavior observations for classroom engagement (see Miller et al., 2015 for a comparison of various screening modalities that could be utilized in future SWTRS validation work). Regardless of the measurement procedures in future studies, analyzing all variables of interest with confirmatory factor analyses (e.g., Iverson & Maguire, 2000) would allow for statistical control and estimation of common method variance (Podsakoff et

al., 2003). Furthermore, structural equation modeling approaches would offer a framework for estimating relations among all variables simultaneously to gain a more sophisticated understanding of their associations than is possible through more traditional multiple regression analyses (Kline, 2016). Additional construct validation work is also called for. Because the results indicated that the WB scales may represent broad meta-constructs that incorporate indicators of other sub-constructs, careful evaluation of these scales in relation to other established measures of student's academic engagement and prosocial behavior is needed.

Even though evidence supporting the FG/FW model of WB was not found as hypothesized, the newly created item pool did nonetheless yield initial support for a structurally valid measure of two important indicators of youth WB at school. A logical next step is replication of these results with a larger, more diverse sample and use of more rigorous analysis methods (e.g., SEM). Beyond this, because the present study only investigated basic science questions related to the structural and concurrent validity of teacher-reported youth WB, future research should extend to test risk classification accuracy as well as the applied utility of these WB scales in various school-based service delivery contexts. Some examples may include examining how AES and SPS scores could be used to inform Tier 1 and 2 interventions or testing the relative sensitivity of the scores to change over time when used as progress monitoring instruments in a schoolwide MTSS for mental health. Having the ability to draw from a combination of both basic science validation and treatment utility evidence is ideal when considering which instruments would be most appropriate for a school's universal mental health screening initiatives.

References

Albers, C. A., & Kettler, R. J. (2014). Best practices in universal screening. In P. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (pp. 121–131). Bethesda, MD: The National Association of School Psychologists.

American Psychiatric Association. (2013). Major Depressive Disorder. In *Diagnostic and statistical manual of mental disorders* (5th ed.; pp. 160–168). Arlington, VA: American Psychiatric Publishing.

Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools, 45,* 369–386. doi:10.1002/pits

Armistead, R. J., & Smallwood, D. L. (2014). The National Association of School Psychologists model for comprehensive and integrated school psychological services. In P. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (pp. 9–23). Bethesda, MD: The National Association of School Psychologists.

Bradshaw, C. P., Buckley, J. A., & Ialongo, N. S. (2008). School-based service utilization among urban children with early onset educational and mental health problems: The squeaky wheel phenomenon. *School Psychology Quarterly, 23,* 169–186. doi:10.1037/1045-3830.23.2.169

Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 99–136). Washington, DC: American Psychological Association.

Burns, B. J., Costello, E. J., Angold, A., Tweed, D., Stangl, D., Farmer, E. M., & Erkanli, A. (1995). Children's mental health service use across service sectors. *Health Affairs, 14,* 147–159. doi:10.1377/hlthaff.14.3.147

Caldarella, P., & Merrell, K. W. (1997). Common dimensions of social skills of children and adolescents: A taxonomy of positive behaviors. *School Psychology Review, 26,* 264–278.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309–319.

Cook, C. R. (2013). Manual: Student Internalizing Behavior Screener and Student Externalizing Behavior Screener. Unpublished manual.

Cook, C. R., Rasetshwane, K. B., Truelson, E., Grant, S., Dart, E. H., Collins, T. A., & Sprague, J. (2011). Development and validation of the Student Internalizing Behavior Screener: Examination of reliability, validity, and classification accuracy. *Assessment for Effective Intervention, 36,* 71–79. doi:10.1177/1534508410390486

Cowen, E. L., & Kilmer R. P. (2002). "Positive Psychology": Some plusses and some open issues. *Journal of Community Psychology, 30,* 449–460. doi:10.1002/jcop.10014

D'Agostino, R. B., Belanger, A., & D'Agostino Jr., R. B. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician, 44*, 316–321.

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed.; Vol. 26). Thousand Oaks, CA: SAGE publications, Inc.

DiPerna, J. C. (2006). Academic enablers and student achievement: Implications for assessment and intervention services in the schools. *Psychology in the Schools, 43,* 7–17. doi:10.1002/pits.20125

Dowdy, E., Furlong, M., Raines, T. C., Bovery, B., Kauffman, B., Kamphaus, R. W., ... & Murdock, J. (2014). Enhancing school-based mental health services with a preventive and promotive approach to universal screening for complete mental health. *Journal of Educational and Psychological Consultation, 25,* 179–197. doi:10.1080/10474412.2014.929951

Dowdy, E., Ritchey, K., & Kamphaus, R. W. (2010). School-based screening: A population-based approach to inform and monitor children's mental health needs. *School Mental Health, 2,* 166–176. doi:10.1007/s12310-010-9036-3

Earhart Jr., J., Jimerson, S. R., Eklund, K., Hart, S. R., Jones, C. N., Dowdy, E., & Renshaw, T. L. (2009). Examining relationships between measures of positive behaviors and negative functioning for elementary school children. *The California School Psychologist, 14*, 97–104.

Eklund, K., & Dowdy, E. (2013). Screening for behavioral and emotional risk versus traditional school identification methods. *School Mental Health, 6,* 40–49. doi:10.1007/s12310-013-9109-1

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*, 121–138. doi: 10.1037/1082-989X.12.2.121

Field, A. (2013). Discovering statistics using IBM SPSS statistics. London, England: SAGE publications Ltd.

Finch, W. H., Bolin, J. E., & Kelley, K. (2014). *Multilevel modeling using R*. Boca Raton, FL: Taylor & Francis.

Fredricks, J., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74,* 59–109. doi:10.3102/00346543074001059

Fredricks, J., McColskey, W., Meli, J., Mordica, J., Montrosse, B., & Mooney, K. (2011). *Measuring student engagement in upper elementary through high school: A description of 21 instruments*. (Issues & Answers Report, REL 2011–No. 098). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from http://ies.ed.gov/ncee/edlabs.

Furlong, M. J., Gilman, R., & Huebner, E. S. (Eds.). (2014). *Handbook of positive psychology in the schools* (2nd ed.). New York, NY: Routledge.

Furlong, M. J., You, S., Renshaw, T. L., O'Malley, M. D., & Rebelez, J. (2013). Preliminary development of the Positive Experiences at School Scale for elementary school children. *Child Indicators Research, 6,* 753–775. doi:10.1007/s12187-013-9193-7

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

Greco, L. A., Lambert, W., & Baer, R. A. (2008). Psychological inflexibility in childhood and adolescence: Development and evaluation of the Avoidance and Fusion Questionnaire for Youth. *Psychological Assessment, 20*, 93–102. doi:10.1037/1040-3590.20.2.93

Greenspoon, P. J., & Saklofske, D. H. (2001). Toward an integration of subjective wellbeing and psychopathology. *Social Indicators Research, 54,* 81–108. doi:10.1023/A:1007219227883

Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1987). The treatment utility of assessment: A functional approach to evaluating assessment quality. *American Psychologist, 42,* 963– 974.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Huta, V. (2014). When to use hierarchical linear modeling. *The Quantitative Methods for Psychology*, *10*(1), 13–28.

Individuals with Disabilities Education Improvement Act, 20 U.S.C. § 1400 *et seq.* (2004).

Iverson, R. D., & Maguire, C. (2000). The relationship between job and life satisfaction: Evidence from a remote mining community. *Human Relations, 53,* 807–839.

Jimerson, S. R., Campos, E., & Greif, J. L. (2003). Toward an understanding of definitions and measures of school engagement and related terms. *California School Psychologist, 8,* 7–27.

Kamphaus, R. W., DiStefano, C., Dowdy, E., Eklund, K., & Dunn, A. R. (2010). Determining the presence of a problem: Comparing two approaches for detecting youth behavioral risk. *School Psychology Review*, *39,* 395–407.

Keyes, C. L. M. (2006). Mental health in adolescence: Is America's youth flourishing? *American Journal of Orthopsychiatry, 76,* 395–402. doi:10.1037/0002-9432.76.3.395

Keyes, C. L. M. (2007). Promoting and protecting mental health as flourishing: A complementary strategy for improving national mental health. *American Psychologist, 62,* 95–108. doi:10.1080/17439760902844228

Keyes, C. L. M., & Annas, J. (2009). Feeling good and functioning well: Distinctive concepts in ancient philosophy and contemporary science. *The Journal of Positive Psychology, 4,* 197–201. doi:10.1080/17439760902844228

Kilgus, S. P., Chafouleas, S. M., & Riley-Tillman, T. C. (2013). Development and initial validation of the Social and Academic Behavior Risk Screener for elementary grades. *School Psychology Quarterly*, *28*, 210–226. doi:10.1037/spq0000024

Kilgus, S. P., Eklund, K., von der Embse, N. P., Taylor, C. N., & Sims, W. A. (2016). Psychometric defensibility of the Social, Academic, and Emotional Behavior Risk Screener (SAEBRS) Teacher Rating Scale and multiple gating procedure within elementary and middle school samples. *Journal of School Psychology*, *58*, 21–39. doi:10.1016/j.jsp.2016.07.001

Kim, E. K., Furlong, M. J., Dowdy, E., & Felix, E. D., (2014). Exploring the relative contributions of the strength and distress components of dual-factor complete mental health screening. *Canadian Journal of School Psychology*. Advance online publication. doi:10.1177/0829573514529567

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Publications, Inc.

Lane, K. L., Little, M. A., Casey, A. M., Lambert, W., Wehby, J., Weisenbach, J. L., & Phillips, A. (2009). A comparison of systematic screening tools for emotional and behavioral disorders. *Journal of Emotional and Behavioral Disorders*, *17*, 93–105. doi:10.1177/1063426608326203

Lane, K. L., Oakes, W., & Menzies, H. (2010). Systematic screenings to prevent the development of learning and behavior problems: Considerations for practitioners, researchers, and policy makers. *Journal of Disability Policy Studies*, *21*, 160–172. doi:10.1177/1044207310379123

Lyons, M. D., Huebner, E. S., & Hills, K. J. (2012). The dual-factor model of mental health: A short-term longitudinal study of school-related outcomes. *Social Indicators Research, 114*, 549–565. doi:10.1007/s11205-012-0161-2

Marmorstein, N. R., Iacono, W. G., & Malone, S. M. (2010). Longitudinal associations between depression and substance dependence from adolescence through early adulthood. *Drug and Alcohol Dependence, 107*, 154–160. doi:10.1016/j.drugalcdep.2009.10.002

Mathyssek, C. M., Olino, T. M., Velhurst, F. C., & van Oort, F. V. A. (2012). Childhood internalizing and externalizing problems predict the onset of clinical panic attacks over adolescence: The TRAILS study. *PLoS ONE, 7*, e51564. doi:10.1371/journal.pone.0051564

Merikangas, K.J., He, J. P., Burstein, M., Sonja, A., Swanson, S.A., Avenevoli, S., ...& Swendsen, J. (2010). Lifetime prevalence of mental disorder in U.S. adolescents: Results from the national comorbidity study—Adolescent supplement (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry, 49*, 980−989. doi:10.1016/ j.jaac.2010.05.017

Merrell, K. W. (2008). *Helping students overcome depression and anxiety: A practical guide* (2nd ed.). New York, NY: Guilford Press.

Miller, F. G., Cohen, D., Chafouleas, S. A., Riley-Tillman, T. C., Welsh, M. E., & Fabiano, G. A. (2015). A comparison of measures to screen for social, emotional, and behavioral risk. *School Psychology Quarterly, 30*, 184–196. doi:10.1037/spq0000085

National Association of School Psychologists (2010). Model for comprehensive and integrated school psychological services. Retrieved from http://www.nasponline.org/standards/2010standards/2_PracticeModel.pdf

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team (2016). nlme: Linear and nonlinear mixed effects models. R package version 3.1-128. Retrieved from http://CRAN.R-project.org/package=nlme.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88,* 879–903. doi:10.1037/0021-9010.88.5.879

Quay, H. C. (1986). Classification. In H. C. Quay & J. S. Werry (Eds.), *Psychopathological disorders of childhood* (3rd ed., pp. 1–34). New York: Wiley.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Rapport, M. D., Denney, C. B., Chung, K.-M., & Hustace, K. (2001). Internalizing behavior problems and scholastic achievement in children: Cognitive and behavioral pathways as mediators of outcome. *Journal of Clinical Child & Adolescent Psychology, 30,* 536–551. doi:10.1207/S15374424JCCP3004_10

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Renshaw, T. L. (2015). A replication of the technical adequacy of the Student Subjective Wellbeing Questionnaire. *Journal of Psychoeducational Assessment,* 1–12. doi:10.1177/0734282915580885

Renshaw, T. L., & Cohen, A. S. (2014). Life satisfaction as a distinguishing indicator of college student functioning: Further validation of the two-continua model of mental health. *Social Indicators Research, 117,* 319–334. doi:10.1007/s11205-013-0342-7

Renshaw, T. L., Furlong, M. J., Dowdy, E., Rebelez, J., Smith, D. C., O'Malley, M. D., . . . Frugård Strøm, I. (2014). Covitality: A synergistic conception of youths' mental health. In M. J. Furlong, R. Gilman & E. S. Huebner (Eds.), *Handbook of positive psychology in the schools* (2nd ed., pp. 12–32). New York, NY: Routledge.

Renshaw, T. L., Long, A. C. J., & Cook, C. R. (2014). Assessing adolescents' positive psychological functioning at school: Development and validation of the Student Subjective Wellbeing Questionnaire. *School Psychology Quarterly*. Advance online publication. doi:10.1037/spq0000088

Rudolph, K. D., & Klein, D. N. (2009). Exploring depressive personality traits in youth: Origins, correlates, and developmental consequences. *Development and Psychopathology, 21,* 1155–1180. doi:10.1017/S0954579409990095

Seligman, M. E. P. (2002). Positive psychology, positive prevention, and positive therapy. In C. R. Snyder & S. J. Lopez (Eds.), *Handbook of positive psychology* (pp. 3–12). New York, NY: Oxford University Press.

Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist*, *55*, 5–14. doi:10.1037//0003-066X.55.1.5

Snijders, T. A. B., & Bosker, R. J. (1999) *Multilevel Analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: SAGE Publications, Inc.

Stoiber, K. C. (2014). A comprehensive framework for multitiered systems of support in school psychology. In P. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (pp. 41–70). Bethesda, MD: The National Association of School Psychologists.

Strein, W., Hoagwood, K., & Cohn, A. (2003). School psychology: A public health perspective I. Prevention, populations, and systems change. *Journal of School Psychology*, *41*, 23–38. doi:10.1016/S0022-4405(02)00142-5

Suldo, S., Thalji, A., & Ferron, J. (2011). Longitudinal academic outcomes predicted by early adolescents' subjective wellbeing, psychopathology, and mental health status yielded from a dual factor model. *The Journal of Positive Psychology, 6*, 17–30. doi:10.1080/17439760.2010.536774

Suldo, S. M. & Shaffer, E. J. (2008). Looking beyond psychopathology: The dual-factor model of mental health in youth. *School Psychology Review*, *37*, 52–68. Retrieved from: https://www.researchgate.net/publication/228656864

Volpe, R. J., DuPaul, G. J., DiPerna, J. C., Jitendra, A. K., Lutz, J. G., Tresco, K., & Junod, R. V. (2006). Attention deficit hyperactivity disorder and scholastic achievement: A model of mediation via academic enablers. *School Psychology Review*, *35*, 47– 61.

von der Embse, N. P., Pendergast, L. L., Kilgus, S. P., & Eklund, K. R. (2015). Evaluating the applied use of a mental health screener: Structural validity of the Social, Academic, and Emotional Behavior Risk Screener. *Psychological Assessment*. Advance online publication. doi:10.1037/pas0000253

Warnes, E. D., Sheridan, S. M., Geske, J., & Warnes, W. A. (2005). A contextual approach to the assessment of social skills: Identifying meaningful behaviors for social competence. *Psychology in the Schools*, *42*, 173–187. doi:10.1002/pits.20052

Whitley, B. E., & Kite, M. E. (2012). Factor analysis, path analysis, and structural equation modeling. In Author (Eds.), *Principles of research in behavioral science* (3rd ed.; pp. 338–361). New York, NY: Routledge.

Wickham, H. (2014). Tidy data. *The Journal of Statistical Software*, *59*, 1–24. Retrieved from: http://www.jstatsoft.org/v59/i10/

World Health Organization. (2004). Promoting mental health: Concepts, emerging evidence, practice (Summary report). Geneva: Author. Retrieved from: http://www.who.int/mental_health/evidence/en/promoting_mhh.pdf

Ysseldyke, J., & Reschly, D. J. (2014). The evolution of school psychology: Origins, contemporary status, and future directions. In P. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (pp. 71–84). Bethesda, MD: The National Association of School Psychologists.

Appendix A
IRB Approval

ACTION ON EXEMPTION APPROVAL REQUEST

Institutional Review Board
Dr. Dennis Landin, Chair
130 David Boyd Hall
Baton Rouge, LA 70803
P: 225.578.8692
F: 225.578.5983
irb@lsu.edu | lsu.edu/irb

TO: Anthony Roberson
Psychology

FROM: Dennis Landin
Chair, Institutional Review Board

DATE: February 8, 2016

RE: **IRB#** E9745

TITLE: Initial Development and Validation of the Student Wellbeing Teacher Report Scale

New Protocol/Modification/Continuation: New Protocol

Review Date: 2/1/2016

Approved X **Disapproved**

Approval Date: 2/8/2016 **Approval Expiration Date:** 2/7/2019

Exemption Category/Paragraph: 1, 2b

Signed Consent Waived?: No

Re-review frequency: (three years unless otherwise stated)

LSU Proposal Number (if applicable):

Protocol Matches Scope of Work in Grant proposal: (if applicable)

By: Dennis Landin, Chairman DLandin

PRINCIPAL INVESTIGATOR: PLEASE READ THE FOLLOWING –
Continuing approval is CONDITIONAL on:

1. Adherence to the approved protocol, familiarity with, and adherence to the ethical standards of the Belmont Report, and LSU's Assurance of Compliance with DHHS regulations for the protection of human subjects*
2. Prior approval of a change in protocol, including revision of the consent documents or an increase in the number of subjects over that approved.
3. Obtaining renewed approval (or submittal of a termination report), prior to the approval expiration date, upon request by the IRB office (irrespective of when the project actually begins); notification of project termination.
4. Retention of documentation of informed consent and study records for at least 3 years after the study ends.
5. Continuing attention to the physical and psychological well-being and informed consent of the individual participants, including notification of new information that might affect consent.
6. A prompt report to the IRB of any adverse event affecting a participant potentially arising from the study.
7. Notification of the IRB of a serious compliance failure.
8. **SPECIAL NOTE: When emailing more than one recipient, make sure you use bcc. Approvals will automatically be closed by the IRB on the expiration date unless the PI requests a continuation.**

**All investigators and support staff have access to copies of the Belmont Report, LSU's Assurance with DHHS, DHHS (45 CFR 46) and FDA regulations governing use of human subjects, and other relevant documents in print in this office or on our World Wide Web site at http://www.lsu.edu/irb*

Appendix B
Teacher Demographic Survey

Teacher Demographic Information

1. **Enter Unique Letter Code**

2. **Grade You Teach**
   Mark only one oval.
   - Kindergarten
   - 1st Grade
   - 2nd Grade
   - 3rd Grade
   - 4th Grade
   - 5th Grade

3. **Gender**
   Mark only one oval.
   - Female
   - Male

4. **Age**

5. **Number of Years Teaching**

6. **Highest Degree Earned**

7. **Race/Ethnicity**
Mark only one oval.

- American Indian/Alaskan Native
- Asian
- Black/African American
- Hispanic/Latino(a)
- Native Hawaiian/Other Pacific Islander
- White/Caucasian
- Two or More Races/Ethnicities

Appendix C
Student Behavior Survey

Teacher-Report Student Behavior Survey

1. **Enter Unique Letter Code**

2. **Student Gender**
   Mark only one oval.
   - ◯ Male
   - ◯ Female

3. **Student Age**

4. **Student Race/Ethnicity**
   Mark only one oval.
   - ◯ American Indian/Alaskan Native
   - ◯ Asian
   - ◯ Black/African American
   - ◯ Hispanic/Latino(a)
   - ◯ Native Hawaiian/Other Pacific Islander
   - ◯ White/Caucasian
   - ◯ Two or More Race/Ethnicities

Student Behavior Frequency Ratings

NOTE: The following two sections use different response options. Please keep this in mind when responding.

5. **For each item, please indicate how frequently the student showed the behavior over the last two months [Never, Rarely, Occasionally, or Frequently/Almost Always]**
Mark only one oval per row.

	Never	Rarely	Occasionally	Frequently/Almost Always
Bullies others				
Defiant or oppositional to adults				
Withdrawn				
Complains about being sick or hurt				
Seems sad or unhappy				
Fights or argues with peers				
Bullied by peers				
Clings to adults				
Has difficulty sitting still				
Nervous, worried, or fearful				
Lies to get out of trouble				
Gets angry easily				
Spends time alone				
Disrupts class activities				

6. **For each item, please indicate how frequently the student showed the behavior over the last two months [Almost Never, Sometimes, Often, or Almost Always]**
Mark only one oval per row.

	Almost Never	Sometimes	Often	Almost Always
Follows directions in class				
Listens to teachers and staff at school				
Classmates like to work or play with him/her				
Needs little supervision when given an assignment				
Inquisitive or interested in learning new things.				
Engaged in learning				
Seems happy during class				
Comfortable working independently on school assignments				
Handles frustrations at school well				
Stays on-task when given an assignment				
Participates meaningfully in class activities				
Alert during lessons and when listening to instructions				
Respectful to classmates				
Enjoys working with other students during group activities				
Shows self-control when frustrated at school.				
Confident when faced with new or				

challenging material.				
Classmates are respectful toward him/her				
Is friendly with classmates				
Respectful to teachers				
Keeps hands/feet to self during class time				
Stays focused during class activities				
Optimistic they will succeed in school				
Gets along well with classmates				
Seems relaxed and at ease during school				
Willing to try new activites at school				
Approachable and easy to get along with at school				
Peaceful during class time				
Shows excitement for class activities				
Plays well with other students				
Sociable with other students during free time				
Remains calm when facing difficult situations at school.				
Laughs at appropriate times in school.				
Treats classmates kindly				
Obeys classroom rules				
Well-behaved during class				
Smiles at school				

Student Outcome Measures

7. **In the past two months, about what percent of time was the student on task during class?**
Mark only one oval.

- 0–10%
- 11–20%
- 21–30%
- 31–40%
- 41–50%
- 51–60%
- 61–70%
- 71–80%
- 81–90%
- 91–100%

8. **In the past two months, how well has the student performed in English Language Arts?**
Mark only one oval.

- ◯ Far below grade level
- ◯ Below grade level
- ◯ At grade level
- ◯ Above grade level
- ◯ Far above grade level

9. **In the past two months, how well has the student performed in Math?**
Mark only one oval.

- ◯ Far below grade level
- ◯ Below grade level
- ◯ At grade level
- ◯ Above grade level
- ◯ Far above grade level

10. **In the past two months, about how many full days of school has the student missed?**

_____

Vita

A native of St. Louis, Missouri, Anthony J. Roberson graduated with honors from Truman State University in 2013. At TSU, he received a Bachelor of Science degree in Psychology and completed minors in Statistical Methods and Music. He has since enrolled in graduate school at Louisiana State University where he is pursuing his Ph.D. in School Psychology under the mentorship of Dr. Tyler Renshaw. His research interests broadly concern youth wellbeing and improving psychological service delivery within school systems.