

A PSYCHOMETRIC INVESTIGATION OF THE UNIVERSAL
BEHAVIOR SCREENER (UBS): A SOCIAL, EMOTIONAL, AND BEHAVIORAL
SCREENER FOR ELEMENTARY SCHOOL STUDENTS

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE
UNIVERSITY OF HAWAI'I AT MĀNOA IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

PSYCHOLOGY

August 2021

By

Kaitlin A. Hill

Dissertation Committee:

Brad Nakamura, Chairperson

Yiyuan Xu

Emily Daubert

David Royer

Min Liu

Keywords: universal screening, measurement, school mental health,
social emotional learning, prevention

Dedication

To everyone who thinks a little differently, who feels a little differently,
who struggles a little differently, who doesn't fit neatly inside the box –
you are not alone, you are worthy, you are valued,
and your voice is needed in this world.

Acknowledgments

I would first like to thank my advisor, Dr. Brad Nakamura, for his guidance and dedication to my development as a clinical psychologist throughout our time together. Over the years, I learned (among many things) the value of humility in my work as a scientist and practitioner and grew my passion for advocating for and serving the children and families in our community. This project would not have been possible without Shane Myers and Dr. Kim McDonald, my partners at the State of Hawai‘i Department of Education, who developed the UBS, provided me with the archival data used for this study, and were so helpful throughout this process. It was a joy to work with and learn from them, and our community is lucky to have such champions of children’s mental health in our schools. I also want to thank my committee for all of their thoughtful feedback and helpful insights on this project.

I have been so fortunate to have such a brilliant, passionate, and collaborative grad school family; I am a better scientist, therapist, supervisor, advocate, and friend because of their friendship and example. I would not be where I am or who I am today without my parents and their truly endless support. There are no words to express the depths of my gratitude. They encouraged me to always dream big and helped me develop the grit I needed to make it happen. I owe so much to my grandparents and great-grandmother, who inspire me daily with their resilience, integrity, and lives lived in service of others. I want to thank Scott for sharing this dream with me and being a true partner in everything. I am so excited for the adventures ahead. Lastly, I want to thank the many mentors who have invested in me over the years, giving their time and wisdom so freely to help me grow. I am so grateful for the ways in which each of them inspired me to dream more, to learn more, and to become more.

Abstract

This study was an initial psychometric investigation of the Universal Behavior Screener (UBS), a nine-item screening measure developed by psychologists within the Hawai'i Department of Education (HIDOE) to identify elementary school youth at risk for social, emotional, and behavioral concerns. Data from nine teachers reporting on 230 students at one HIDOE elementary school were used to examine the UBS factor structure, reliability, convergent validity, and concurrent and predictive criterion-related validity in relation to the Behavior Intervention Monitoring Assessment System, 2 (BIMAS-2). Results from exploratory factor analysis supported a two-factor structure for the UBS (i.e., Social/Emotional Engagement and Academic Readiness). Reliability analyses suggested adequate internal consistency and test-retest reliability of UBS Total and subscale scores. Findings from validity analyses were mixed. Significant correlations were found between UBS and BIMAS-2 subscales in expected patterns. However, both UBS subscales related most strongly to BIMAS-2 Cognitive/Attention and Academic Functioning subscales, contrary to hypotheses. Concurrent criterion-related validity analyses found stronger sensitivity values for hypothesized UBS and BIMAS-2 subscale comparisons; however, no UBS subscale was found to reach an adequate level of sensitivity (i.e., > 75%) in classifying students scoring in the at-risk range on any BIMAS-2 subscale. Sensitivity estimates were considerably worse for predicting short-term future BIMAS-2 subscale at-risk status compared to concurrent validity values. However, specificity and negative predictive values for all UBS and BIMAS-2 subscale comparisons were similarly high (i.e., > .75) for both concurrent and predictive analyses. Implications of findings from this initial examination of the UBS are discussed along with directions for further psychometric investigations of the measure.

Table of Contents

Acknowledgments.....	iii
Abstract.....	iv
List of Tables	viii
List of Figures.....	x
Introduction.....	1
Universal Screening	2
Schoolwide Positive Behavioral Interventions and Supports	4
Social and Emotional Learning.....	6
Approaches to Universal Screening.....	9
Appropriateness for Intended use	11
Technical Adequacy.....	12
Usability.....	15
State of Hawai‘i Prevention and Early Intervention Efforts	18
The Current Investigation	20
Method.....	28
Participants.....	28
Teachers	28
Students.....	29
Power	31
Measures	33
BIMAS-2.....	33
UBS.....	39
Procedures	42
Analytic Strategy.....	43
Data Diagnostics and Preparation.....	43
Inclusion Criteria for Study Aims.....	46
Calculating BIMAS-2 Risk Status Categories for Aims 3 and 4.....	47
Examining the Influence of Gender on BIMAS-2 Scores	48
Proposed Analyses	48

Aim 1: Examine UBS Factor Structure	48
Aim 2: Assess UBS Reliability.....	50
Aim 3: Investigate UBS Convergent and Concurrent Criterion-Related Validity	51
Aim 4: Explore UBS Short-term Predictive Criterion-Related Validity	55
Results.....	57
Data Diagnostics and Preparation	57
Aim 1: Data Diagnostics and Preparation.....	59
Aims 2 Through 4: Data Diagnostics and Preparation	61
Examining the Influence of Gender on BIMAS-2 Scores	65
Aim 1: UBS Factor Structure.....	66
Aim 2: UBS Reliability	77
Cronbach’s Alpha Coefficients.....	78
Test-Retest Correlation Coefficients.....	78
Aim 3: UBS Convergent and Concurrent Criterion-related Validity.....	80
Convergent Validity.....	82
Concurrent Criterion-Related Validity	88
Aim 4: UBS Short-Term Predictive Criterion-Related Validity.....	94
Short-Term Predictive Validity: Correlations.....	95
Short-Term Predictive Validity: Logistic Regressions	98
Discussion	106
Major Findings	107
Aim 1: UBS Factor Structure.....	107
Aim 2: UBS Reliability.....	108
Aim 3: UBS Convergent and Concurrent Criterion-Related Validity	109
Aim 4: UBS Short-Term Predictive Criterion-Related Validity.....	119
Limitations and Future Directions	120
Study Implications	126
Appendix A: BIMAS-2, Teacher Standard Form	128
Appendix B: BIMAS-2 Subscale Legend.....	129
Appendix C: Item-Level Statistics Table for BIMAS-2 Q2 and Q4	130

Appendix D: Universal Behavior Screener.....	132
Appendix E: Item-Level Statistics Tables for UBS Q1 – Q4	133
Appendix F: Distribution and Scope of UBS & BIMAS-2 Outliers Using z-Scores	136
Appendix G: Aim 3 Logistic Regression and Classification Tables: Q2 UBS Social/Emotional Engagement Subscale & Q2 BIMAS-2 Subscales.....	139
Appendix H: Aim 3 Logistic Regression and Classification Tables: Q2 UBS Academic Readiness Subscale and Q2 BIMAS-2 Subscales	144
Appendix I: Aim 3 Logistic Regression and Classification Tables: Q2 UBS Total and Q2 BIMAS-2 Subscales	149
Appendix J: Aim 4 Logistic Regression and Classification Tables: Q1 UBS Social/Emotional Engagement Subscale and Q4 BIMAS-2 Subscale	154
Appendix K: Aim 4 Logistic Regression and Classification Tables: Q1 UBS Academic Readiness Subscale and Q4 BIMAS-2 Subscales.....	160
Appendix L: Aim 4 Logistic Regression and Classification Tables: Q1 UBS Total and Q4 BIMAS-2 Subscales	166
References.....	172

List of Tables

Table 1. Distribution of Teachers and Students by Grade	28
Table 2. Student Demographic Information	30
Table 3. Reliability Coefficients for BIMAS-2 Subscales at Q2 and Q4	39
Table 4. Q2 UBS Item-Level, Corrected Item-Total Correlations, and Cronbach's Alpha if Item Deleted Statistics.....	60
Table 5. UBS Item-Level Means and Standard Deviations for Q1 – Q4	61
Table 6. Descriptive and Normality Statistics for Q1 – Q4 UBS Total and Subscales.	63
Table 7. Descriptive and Normality Statistics for Q2 and Q4 BIMAS-2 Subscales	64
Table 8. Pearson Bivariate Correlations Between UBS Items 1 – 9.....	67
Table 9. Exploratory Factor Analysis: Factor Loadings for UBS Two-Factor Model	72
Table 10. Cronbach's Alpha Coefficients for UBS Total and Subscale Scores Q1 – Q4	79
Table 11. UBS Social/Emotional Engagement: Test-Retest Bivariate Correlations Q1 – Q4	79
Table 12. UBS Academic Readiness Subscale: Test-Retest Bivariate Correlations Q1 – Q4	80
Table 13. UBS Total: Test-Retest Bivariate Correlations Q1 – Q4.....	80
Table 14. Pearson Bivariate Correlations Between Q2 UBS and Q2 BIMAS-2	83
Table 15. Fisher's z -Tests of Q2 BIMAS-2 Correlations with Each UBS Subscale at Q2	85
Table 16. Fisher's z -Tests Comparing UBS Social/Emotional Engagement Correlations with Hypothesized Versus Non-Hypothesized BIMAS-2 Subscales	87
Table 17. Aim 3 Concurrent Validity Values: Mean Q2 UBS Scores by Q2 BIMAS-2 Risk/Concern Status and Classification Accuracy Estimates	91
Table 18. Pearson Bivariate Correlations Between Q1 UBS and Q4 BIMAS-2	96

Table 19. Fisher's z -Tests of Q4 BIMAS-2 Correlations with Each Q1 UBS Subscale	97
Table 20. Aim 4 Predictive Validity Values: Mean Q1 UBS Scores by Q4 BIMAS-2 Risk/Concern Status and Classification Accuracy Estimates	102

List of Figures

Figure 1. Equations for Calculating Sensitivity, Specificity, and Predictive Values	53
Figure 2. Parallel Analysis Scree Plot.....	68
Figure 3. Scree Plot Produced by Factor Analysis	70

A Psychometric Investigation of the Universal Behavior Screener (UBS):

A Social, Emotional, and Behavioral Screener for Elementary School Students

Over the past two decades, there has been increased pressure on schools to meet the diverse learning needs of students through the incorporation of social, emotional, and behavioral prevention approaches within educational efforts (Every Student Succeeds Act [ESSA], 2015; U.S. Department of Health and Human Services, 2000). This emphasis on school mental health treatment and prevention is rooted in a growing consensus that youths' mental health needs are not being adequately addressed through traditional identification and service delivery models (e.g., parent- or physician-referral to mental health provider, IDEA [Individuals with Disabilities Education Act, 2004] accommodations). In terms of identification, approximately 20% of youth have or will meet criteria for a mental health disorder; however, only a small percentage (i.e., <1% to 16%) of these youth are likely to be identified through typical gate-keeping methods such as pediatrician referral (Briggs-Gowan et al., 2000; Horwitz et al., 1992), and less than 1% will meet the academic impairment criteria necessary to receive school mental health (SMH) services under the IDEA category of emotional disturbance (Lane et al., 2010). For the large portion of students whose mental health concerns go untreated, research indicates they are at increased risk for poor academic achievement, school failure or dropout, and higher rates of missed school, as well as future risks of more severe mental health, delinquency, and conduct problems (Wagner et al., 2005; Zigmond, 2006). The prognosis is still disheartening for the less than 1% of students who do receive services under the traditional *wait to fail* model. The term *wait to fail* relates to the problematic practice of refer-test-place, where only struggling students are referred for an assessment of needs to determine eligibility for services, often only occurring once a student has exhibited significant enough failure or impairment to demand teacher

attention (Cash & Nealis, 2004; Dvorsky et al., 2014). By the time students' behaviors reach the threshold for teacher referral, they may be unlikely to show improvement in academic deficits (Greenbaum et al., 1996; Nelson et al., 2004) and over time, are at risk for underemployment, poor community adjustment, and substance abuse, among other sequelae (Bullis & Yovanoff, 2006; Wagner & Davis, 2006; Zigmond, 2006).

Taken together, these findings highlight the need for prevention and early intervention approaches that can better identify youth who might otherwise fall through the cracks and subsequently develop academic deficits or other negative trajectories. Federal and state initiatives in the last two decades have reflected this sentiment, with calls for increased focus on school-based mental health services, the inclusion of social and emotional learning, and prevention and early identification of social, emotional, and behavioral challenges (ESSA, 2015; President's New Freedom Commission on Mental Health, 2003; U.S. Public Health Service, 2000). Integral to any prevention or early intervention initiative, however, is the accurate identification of youth who could benefit from such supports, as well as monitoring the effectiveness of supports provided. Given the poor outcomes typically associated with the *wait to fail* model, newer initiatives have focused on incorporating universal screening procedures to identify students for services before problems escalate.

Universal Screening

Universal screening refers to procedures conducted with all students regardless of whether they exhibit any known risk factors (Levitt et al., 2007). This is differentiated from other screening efforts, such as selected and indicated screening, which are conducted with students who have elevated risk for or have been diagnosed with a mental health problem, respectively (Levitt et al., 2007). The goal of universal screening is to “identify childhood problems before

the behaviors exceed the threshold for parent or teacher referral for services" (Dvorsky et al., 2014, p. 298; Severson et al., 2007). Traditionally, approaches to early identification for youth have relied primarily on adult referral, which has been found to be insufficient, as these referrals tend to lead to identifying only the children who are already exhibiting failure or impairment (Dvorsky et al., 2014; Lane et al., 2010). Consequently, students exhibiting emerging signs of emotional or behavioral concerns and who are non-disruptive (e.g., students with anxiety or depression) may be missed (Albers et al., 2007). This runs counter to the purpose of early identification, which seeks to catch students at the initial onset of behavioral or emotional concerns, or even earlier at the first signs of elevated risk for the development of those concerns. Accurate and comprehensive early identification is imperative because the earlier a student is identified, the earlier they can receive treatment, and earlier treatment tends to mean greater chance of success (Lane et al., 2010). Another benefit to identifying problems before they escalate is that doing so allows interventions to be put in place at a less intensive level that is more likely to be within the expertise, budget, and resources of school staff compared to treatment for more severe problems (Dowdy et al., 2010; Lane et al., 2010). Universal screening is often used by schools within a broader approach to prevention and intervention whereby students are first identified as not at risk, at risk, or currently exhibiting signs of concerns, and then matched to appropriate services. Many of these prevention models involve multi-tiered supports, incorporating three levels of preventive supports: universal, selected, and indicated (The Collaborative for Academic, Social, and Emotional Learning [CASEL], 2018; Lannie et al., 2010; Weist et al., 2014). To provide context for the way in which universal screening approaches may be utilized within larger prevention efforts, I will first review two common prevention approaches in the literature, one multi-tiered framework focused on behavior (i.e.,

schoolwide positive behavioral interventions and supports), and the other, an approach to social and emotional functioning (i.e., social and emotional learning curricula).

Schoolwide Positive Behavioral Interventions and Supports

One of the most widely used school-based prevention frameworks for behavioral concerns is schoolwide positive behavioral interventions and supports (SWPBIS; Lannie et al., 2010; Lewis & Sugai, 1999; Sugai & Horner, 2006; see also Bradshaw et al., 2015). SWPBIS can be used alone, or within larger and more comprehensive models of prevention and intervention, such as the comprehensive, integrated, three-tiered (Ci3T) model of prevention (Lane et al., 2010). Within the Ci3T model, SWPBIS is used alongside other prevention approaches focused on social (i.e., validated social and emotional learning [SEL] curricula) and academic (i.e., validated academic curricula and instructional standards) to support students holistically. Some of the core tenets of SWPBIS include a three-tiered model of supports increasing in intensity, schoolwide behavioral systems and procedures taught and reinforced by all adults, and data-based decision-making to guide all components of implementation (Lane et al., 2009). The SWPBIS framework facilitates prevention of the poor outcomes sometimes evidenced in the literature for students under traditional *wait to fail* approaches by providing behavioral supports to all students in a scaffolding manner (i.e., tiers of support) based on student need at the first sign of concern. Specifically, each tier offers increasingly intensive interventions to meet students' needs, and high-quality data are used for decisions about which supports are appropriate for meeting students' needs at each tier (CASEL, 2018).

The least intensive level, Tier 1 (i.e., universal prevention, approximately 80% or higher of the student population; Lane et al., 2009; Sugai & Horner, 2006) is aimed at all students and focuses on creating a socially predictable, safe, consistent, and positive culture that allows the

maximization of teaching and learning opportunities (Horner et al., 2009). At this universal level, behavioral systems are established schoolwide to teach, prompt, and reinforce students' use of desired and appropriate behaviors in school in order to (a) increase use of prosocial behaviors in place of problem behaviors, (b) utilize adult attention and reinforcement to create a safer and more respectful school environment, (c) make learning environments more preventive, positive, and predictable, and (d) identify and strategically support students who exhibit more resistant problem behavior (Bradshaw et al., 2014). The remaining 20% of students are identified as likely needing additional behavioral supports. Approximately 10-15% of the student population will require Tier 2 supports (i.e., selective supports; Lane et al., 2009) while around 5-10% of students will need Tier 3 supports (i.e., individualized or indicated supports; Lane et al., 2009). Tier 2 addresses the needs of students who are at risk of developing behavior or mental health concerns (Bradshaw et al., 2014) and often includes supplemental small group behavioral skill building (e.g., anger management, social anxiety groups) or brief individualized interventions conducted in the general education classroom (e.g., daily report cards, planner checks; Dvorsky et al., 2014). Tier 3 is the most intensive level of supports and focuses on addressing the needs of students displaying early to significant signs of behavioral and/or mental health problems (Bradshaw et al., 2014) through individualized interventions often involving specialized services (e.g., school- or community-based mental health services).

High fidelity SWPBIS implementation at the elementary school level has evidenced improvements across various domains of student outcomes, including academic performance (Bradshaw et al., 2010; Horner et al., 2009; McIntosh et al., 2011; Nelson et al., 2002); disruptive behavior and concentration problems, (Bradshaw et al., 2010; Bradshaw et al., 2012; McCurdy et al., 2003; Nelson et al., 2002); bullying and peer rejection (Waasdorp et al., 2012);

and social and emotional functioning (Waasdorp et al., 2012). Other benefits of SWPBIS have been shown for school climate (Horner et al., 2009) and staff, including improved self-efficacy (Kelm & McIntosh, 2012; Ross & Horner, 2006; Ross et al., 2012) and reduced burnout (Ross et al., 2012). However, the effectiveness of SWPBIS and any multi-tiered prevention approach is largely dependent on the ability to accurately identify and match students to available services and supports.

Social and Emotional Learning

While SWPBIS is focused primarily on the behavioral domain (Lane et al., 2020), recommendations in the prevention literature focus not only on behavioral management, but social and emotional supports as well. One strategy for promoting students' social and emotional well-being that has gained popularity in research and practice nationwide is social and emotional learning (SEL; Durlak et al., 2011; Zins & Elias, 2006). According to Zins and Elias (2006), SEL is generally a curriculum-based approach aimed at improving students' emotional awareness and management, effective problem solving, and relationships with others through targeting a combination of thoughts, behaviors, and emotions. Social and emotional skills are seen as integral to the learning process, as learning is inherently a social and interactive process (CASEL, 2018). Thus, social and emotional deficits are expected to manifest problematically in the classroom environment, disrupting the performance of the student and potentially their classmates as well (Gresham, 2015). For this reason, it is common for schools to focus prevention efforts on the social and emotional development of their students (Weissberg et al., 2015). The term SEL is used to reflect the notion that social/emotional and life skills must be taught explicitly (i.e., learned) in the same way as academic skills (e.g., reading or math) for them to be internalized and integrated into the child's lifelong repertoire (Elias, 2006). In this

way, there is a shift from external control of a child's behavior to the child's behavior being guided by internally based values and beliefs, which have developed through mastery of social and emotional competencies (Durlak et al., 2011).

CASEL has identified five interrelated cognitive, affective, and behavioral competencies as a framework for SEL (Zins et al., 2007), generally aligning with Denham's (2005) unifying model of social and emotional skills. According to Zins and colleagues (2007) and Denham (2005), these five interrelated competencies include three emotional competence skills: (a) self-awareness (i.e., ability to accurately recognize one's emotions, thoughts, and behavior and understand their interconnectedness), (b) self-management (i.e., ability to regulate one's emotions, thoughts, and behaviors effectively), (c) social awareness (i.e., ability to take the perspectives of others and to empathize and feel compassion), and two relational/prosocial skills: (d) relationship skills (i.e., ability to establish and maintain healthy relationships and act in alignment with social norms), and (e) responsible decision-making or social problem solving (i.e., ability to make constructive and respectful choices about personal behavior and social interactions across settings). The overarching conceptual model of SEL proposes that strengthening student's competencies in core areas (e.g., emotional regulation, self-management) influences academic performance both directly (e.g., more time engaged in instruction due to better regulation of internal resources for goal-directed learning) and indirectly (e.g., more effective learning interactions due to increased positive school experiences and feelings of school connectedness; Valiente et al., 2012; Panayiotou et al., 2019). Along with improving the effectiveness of learning interactions, these models posit that such core competencies also reduce known academic and lifelong risk factors (e.g., problem behaviors, emotional stress), which together lead to improved academic performance for these youth in comparison to those with

poor social and emotional skills (CASEL, 2015; Panayiotou et al., 2019). Consistent with these theoretical models, well-implemented and validated SEL programs have been shown to lead to improved academic performance (Corcoran et al., 2018; Durlak et al., 2011; Substance Abuse and Mental Health Services Administration [SAMHSA], 2002); reduced dropout and improved attendance (Wilson et al., 2001); enhanced student behavioral adjustment (e.g., increased prosocial behaviors, reduced conduct and internalizing problems), and improved school attitudes and social and emotional competencies (Durlak et al., 2011).

Like SWPBIS, there is no standard prescription for how schools should implement SEL interventions. Commonly, SEL can be viewed as a universal support for all students and is provided as a curriculum incorporated into general instruction and reinforced throughout schoolwide experiences (e.g., utilized in assemblies, reinforced by all staff). Schools can choose from many different SEL curricula based on leadership priorities and student needs identified through universal screening procedures. These curricula often align with CASEL's (2015) SEL competencies but may have varying levels of empirical support (for reviews of curricula see Corcoran et al., 2018; Rimm-Kaufman & Hulleman, 2015; U.S. Department of Education, 2019). Although SEL can be utilized as a standalone school-based prevention support, many SMH experts advocate for its integration into three-tiered support frameworks, such as SWPBIS, to comprehensively address social, emotional, and behavioral problems, as well as integrate systematic screenings within one comprehensive system (Bear et al., 2015; Bradshaw et al., 2014; Lane et al., 2010; Weissberg et al., 2015). Furthermore, stakeholders in this area also recommend using SEL-focused screening three (i.e., fall, winter, spring) or four (i.e., quarterly) times a year for routine monitoring at the universal level (Bear et al., 2015; Elias, 2006). Studies have found that the comprehensive approach described above can produce superior benefits in

overall youth mental health and reductions in externalizing behaviors compared to SWPBIS or SEL alone (Cook et al., 2015; Durlak et al., 2011).

Approaches to Universal Screening

Regardless of specific prevention approach or combination of approaches (e.g., SEL within the SWPBIS framework), universal screening remains an important component for increasing the effectiveness of any prevention program and ensuring fewer students falling through the cracks. There are numerous options available for universal screeners and it is up to each school to decide which procedures they should implement. These decisions are considered within the context of several factors unique to the school, such as their needs and purpose of screening (e.g., identifying schoolwide needs for SEL programming decisions, identifying specific students needing social, emotional, and behavioral supports within a multi-tiered system), resources to implement the screening protocol, and capacity to address the needs of identified youth through established systems (e.g., SWPBIS; Levitt et al., 2007). While specific considerations are unique to each school, general recommendations are available to help guide screener selection for schools. What follows is a review of the literature on data sources used in universal screening approaches and considerations for selecting a protocol for identifying social, emotional, and/or behavioral concerns in elementary school youth.

Within the school-based prevention literature, several different data sources have been identified for use in assessing student needs for supports and subsequent response to intervention. These sources often include office discipline referrals (ODRs), academic statewide assessments or report card data, attendance records, suspension and expulsion rates, and counseling referrals (Lane et al., 2009; Lannie et al., 2010). Often, combinations of these different data sources are utilized within SWPBIS programs for decisions about which students

require Tier 2 or Tier 3 supports, allowing schools to capitalize on pre-existing data sources for decision-making (Lane et al., 2010). ODRs have received considerable attention in the literature as one of the more useful and relevant tools for identifying students at risk of or exhibiting social, emotional, and/or behavioral concerns and making service-related decisions (Severson et al., 2007; Lane et al., 2009; Lane et al., 2010). Indeed, some research has found links between numbers of referrals, which can be monitored against national or local norms through online data management programs, and a variety of student risk indicators (Severson et al., 2007; Sugai et al., 2000; Lane et al., 2010). However, despite these positive attributes, ODRs alone are not a sufficient means for identification, given their interdependency on teachers' behaviors and varying decision-making thresholds (Lane et al., 2010). Furthermore, ODRs and other common data sources (e.g., suspension rates) have been shown to produce similar rates of identification as teacher referral, as they, too, are unlikely to pick up on soft signs of externalizing concerns (Miller et al., 2015), and are a particularly poor predictor of internalizing concerns (Lane et al., 2010).

Thus, it is strongly recommended schools also utilize systematic screening tools and procedures to increase the accuracy and breadth of early identification efforts (Lane et al., 2009; Lane et al., 2010). Within the literature on universal screening, several published guidelines and reviews have identified core criteria important in evaluating and selecting an optimal systematic screening tool (Dvorsky et al., 2014; Glover & Albers, 2007; Lane et al., 2009; Levitt et al., 2007; Severson et al., 2007). Across guidelines, most of these considerations can be categorized into three broad domains: (a) appropriateness for intended use, (b) technical adequacy, and (c) usability (Dvorsky et al., 2014; Glover and Albers, 2007; Levitt et al., 2007; Sicheloff et al., 2017). To make decisions about screener selection across any of the recommended criteria, the

school must first be clear on their objectives for the screening and their available resources for implementing the screening procedures. Doing so creates the context in which all subsequent decisions are considered. Clearly articulating screening objectives and available resources also helps to increase the likelihood of judicious evaluation of screeners' psychometric properties and ultimately meeting schools' needs (Siceloff et al., 2017). For example, if a school identifies a major concern for their students related to social and emotional skill deficits, the screening tools and procedures they choose should be selected based on their ability to accurately and reliably screen for those key concerns.

Appropriateness for Intended use. Once the school has a clear view of their objective and priorities, all potential screeners can be judged in terms of their appropriateness for use (i.e., ability to meet the school's objectives and priorities for the screening) and their appropriate fit to the school context (Dvorsky et al., 2014; Glover & Albers, 2007). According to Glover and Albers (2007), this includes considerations such as the (a) compatibility with local service delivery needs (e.g., appropriateness of timing and frequency of screening, relevance of identification outcomes), (b) theoretical and empirical support of the screener format and content, which should align with local needs, (c) alignment of screener items with constructs of interest (e.g., if the screener is intended to identify youth with depression, then items should target known symptoms and risk factors for depression), and (d) adequate fit for the population of interest (e.g., contextual and developmental appropriateness of the screener for age range, school setting, diverse students, intended respondents). Along these lines, for schools focused on identifying students with deficits in social and emotional skills, a valid and reliable SEL screening measure would likely be appropriate (Denham, 2015). In contrast, this type of measure would not be appropriate for a school that is interested only in identifying students at risk for

depression or suicide. In terms of domain of interest, broad-based measures (i.e., assessing multiple domains) are typically recommended for universal screening procedures over specific tools (i.e., assessing only one domain) due to the priority of obtaining a time-efficient snapshot across a wide range of difficulties students might be exhibiting (Dowdy et al., 2010; Dvorsky et al., 2014). Furthermore, rating scales are generally viewed by schools as more appropriate for use in schoolwide screening efforts compared to more laborious approaches (e.g., direct observation, interviews) due to their increased efficiency, cost-effectiveness, and versatility (Denham, 2015; Elliott et al., 2015; Humphrey et al., 2011).

Technical Adequacy. Technical adequacy refers to the psychometric properties of the measure, or the ability of the measure to reliably and accurately assess the constructs of interest related to the screening goals (Siceloff et al., 2017). The two major parameters within this domain are the reliability and validity of scores on the screening measure. Reliability refers to the extent to which test scores are dependable and consistent over repeated administrations within a given context and population (American Educational Research Association [AERA] et al., 2014). Common elements of reliability that should be demonstrated by any systematic screener include (a) high internal consistency (i.e., Cronbach alpha of .70 or higher, Lane et al., 2009) indicating that items on a measure or its subscales are consistently measuring a particular domain of interest and (b) strong test-retest stability to ensure the consistency in scores over repeated administrations (Glover & Albers, 2007). Validity refers to the extent to which the interpretation of test scores is supported by theory and evidence for a given use of the test (AERA et al., 2014). Depending on the purpose of the screening measure (e.g., identify youth who currently exhibit concerns and/or predict youth who are likely to develop the problem), evaluations of validity are particularly focused on concurrent and/or predictive validity, both of

which are forms of criterion-related validity. Concurrent validity refers to the extent to which the screener can accurately identify individuals who are currently experiencing difficulties and typically involves examining the relationship between the screener and some criterion measure administered at the same time (Glover & Albers, 2007). Predictive validity is concerned with the extent to which the screener can accurately identify individuals who will and who will not have subsequent behavioral or emotional difficulties (Glover & Albers, 2007). Gold standard references for psychometric evaluations of screening measures note the utility of examining indices such as sensitivity, specificity, positive predictive values (PPV), and negative predictive values (NPV) in assessing the predictive validity and/or concurrent validity of a measure (Glover & Albers, 2007; see also Dvorsky et al., 2014; Levitt et al., 2007; Sicheloff et al., 2017). These indices are defined below as they relate to concurrent validity (i.e., accurately classifying students as having or not having a given condition as determined by some well-validated measure of that condition). For predictive validity, these metrics would be focused on the accurate prediction of students who at some later point go on to develop a given condition. Used here and throughout the paper, the term *condition* refers to the presence of social, emotional, and/or behavioral challenges at the level of needing Tier 2 or Tier 3 supports.

Sensitivity is calculated using only the subsample of students known to have the condition and represents the probability that scores on a screening measure will accurately identify those who indeed have the condition (i.e., identifying true positives), while simultaneously avoiding false negatives (i.e., mis-categorizing a student as *not* having the condition; Trevethan, 2017). PPV is calculated using only the subsample of students testing *positive* on the screening test and indicates the probability of a screening measure correctly identifying the students who actually do have the condition (i.e., identifying true positives) out of

all possible students, while simultaneously avoiding false positives (i.e., mis-categorizing a student as having the condition; Trevethan, 2017). Specificity is calculated using only the subsample of students known to *not* have the target condition and represents the ability of a screening measure to detect a true negative (i.e., students who in fact do *not* have the condition), while at the same time, avoiding false positives. Lastly, NPV is calculated using only the subsample of students testing *negative* on the screening test and is the probability that a screening measure correctly identifies, out of all possible students, those who indeed do *not* have the condition (i.e., identifying true negatives), while simultaneously avoiding false negatives; Trevethan, 2017). Said another way, sensitivity and specificity indicate the degree to which the results on a given measure accurately correspond to the presence or absence of a condition, as determined by results on a reference standard (i.e., a well-validated measure of the condition). In contrast, predictive values are concerned with the effectiveness of the screening measure for accurately categorizing students as having or not having a given condition based on some known variable or reference standard.

Determinations of whether the values found for these indices are adequate for a given screening measure will depend on the purpose of the measure and the consequences of over- or under-identification. For example, the lower the PPV, the greater the chance students are being over-identified as at-risk for a given condition (i.e., false positives), whereas lower sensitivity values result in greater chances of students being under-identified as at risk for a given condition (i.e., false negatives). Each school's priorities will determine which indices they wish to prioritize, and the screening measure selected should align with those priorities. In general, the chance of over-identifying students should be balanced against any potential negative effects of misidentification, such as increased stress on students and families, experiences of

stigmatization, missed opportunities for learning due to subsequent assessments or services, and overuse of programming resources (Glover & Albers, 2007; Levitt et al., 2007). When it comes to mental health screening in schools, however, such costs are typically preferred over the consequences of under-identification, which may result in students missing the opportunity to receive critical supports that could help reduce their risk of experiencing a host of negative academic and life outcomes, including suicide. Additionally, when it comes to internalizing concerns (e.g., depression, anxiety), the consequences of missing a student on a screening measure are particularly concerning as these “silent” disorders are less likely to be identified through other commonly used referral methods (e.g., teacher referral, ODRs, suspensions; Levitt et al., 2007). Therefore, sensitivity is often prioritized at the first gate within multi-gate screening procedures to ensure no students are overlooked who are potentially at risk for concerns, with greater PPV and specificity prioritized during later gates (Glover & Albers, 2007).

Taken together, these different technical adequacy characteristics can provide evidence a measure is assessing the target construct(s) as designed, scores can be interpreted for the intended purpose, and the measure performs in a consistent and dependable manner. These characteristics are important for any measure, but particularly so when scores will be used to make potentially high stakes decisions, such as those often involved in universal screening (e.g., determining access to services).

Usability. While technical adequacy is a necessary consideration in the selection of a systematic screener, even the most precise measure cannot be beneficial for identifying youth if it is not used. The first two considerations help answer the questions: *Is the screener appropriate and does it work for the purpose we have in mind?* The parameter of usability, however, addresses the question, *is the screener likely to be successfully and sustainably implemented?*

Within a schoolwide implementation framework there are numerous stakeholders who have different training backgrounds, different professional responsibilities, and limited time for additional procedures and paperwork. Therefore, it is not surprising implementation decisions may ultimately come down to this question of usability. For a screener to be considered usable, it must demonstrate acceptability, feasibility, and utility for the specified context and purpose. According to Sicheloff and colleagues (2017), *acceptability* means stakeholders (e.g., school administration, respondents) should see the value of the mental health screening approach in helping accomplish the broader goal of promoting children's mental health. Furthermore, for a screening approach to be deemed acceptable, stakeholders should feel the benefits of implementing the procedure outweigh the costs in terms of the time, financial cost, intrusiveness, human resources, and overall additional burden required for implementation (Glover & Albers, 2007).

Feasibility, on the other hand, refers to the extent to which proposed procedures for implementing the universal screener can be implemented and are seen as practical and satisfactory by relevant stakeholders (Sicheloff et al., 2017; see also Levitt et al., 2007). This includes everything from the ease of completing, scoring, and interpreting the screener to the time and cost required for implementation. Lane and colleagues (2009) note that a feasible screener is one that is (a) inexpensive or free and (b) requires minimal time to administer, score, and interpret. If the screening procedure is too resource intensive or financially burdensome, it is unlikely to be sustainable from a school budgeting perspective. Furthermore, even the most precise and inexpensive of screeners is unlikely to be implemented successfully if the primary informant sees the screening protocol as too costly of their time, or if it is seen as too cumbersome by those who are tasked with scoring, interpreting, or integrating into decision-

making (Lane et al., 2010; Wigelsworth et al., 2010). Issues related to informant perspectives on screening are especially important, given this can directly impact the validity of the results obtained from the measure due to issues such as fatigue or frustration (Dvorsky et al., 2014). Often, teachers serve as the primary informants for universal screening procedures, as they have the most interaction with students and have a large body of experiences from which to draw comparisons about appropriate developmental expectations (Wigelsworth et al., 2010). Given the large class sizes typical in American public schools, the total administration time across all students in the class must be considered, not just the completion time for a single student. Thus, it is understandable that even shorter measures could be seen as overly burdensome within the context of the large number of responsibilities already placed on teachers.

The final usability consideration is the *utility* of the measure, or how useful the results of the screener are to stakeholders, particularly related to helping the school achieve their identified objective (e.g., decreasing office referrals, improving social and emotional skills, increasing inclusion rates; Glover & Albers, 2007). To achieve this, the universal screener must be situated within a larger continuum of programs and services that can address student needs as they are identified (Glover & Albers, 2007; Sicheloff et al., 2017). Ultimately, the screener should be seen as a beneficial tool for the school in decision-making. As Sicheloff and colleagues (2017) add, the screener should provide important baseline data on students for use in progress monitoring.

Overall, experts in the field recommend that measures used for universal screening should not only demonstrate technical adequacy, but be cost efficient, require minimal additional effort, align with the core values and aims of the school, and help address a problem identified as a priority by stakeholders in the school (Severson et al., 2007, see also Glover & Albers, 2007; Levitt et al., 2007; Sicheloff et al., 2017). This large number of competing priorities and

considerations can make the act of selecting a screener a complicated and resource intensive process. Furthermore, there may be external constraints imposed on stakeholders influencing their priorities and impacting the way schools weigh these different considerations (e.g., financial benefits such as discounts for statewide contracts with a publisher for a given screening system, leadership preference for state- or district-wide consistency in the screening measure utilized). While it is important to acknowledge the complex factors facing schools when selecting a screening measure, it is still paramount that the tool chosen by the school is psychometrically sound for stakeholders' intended use.

State of Hawai'i Prevention and Early Intervention Efforts

The current study focuses on the initial evaluation of a novel universal screener (i.e., Universal Behavior Screener; UBS) developed by school and clinical psychologists in the Hawai'i Department of Education (HIDOE) to provide local schools a more feasible and acceptable cost-free alternative to the monitoring system purchased by the state, the Behavior Intervention Monitoring Assessment System (BIMAS-2; McDougal et al., 2011). The BIMAS-2 is an assessment tool and monitoring system designed for the screening of social, emotional, behavioral, and academic difficulties, as well as ongoing progress monitoring (McDougal et al., 2011). The BIMAS-2 is considered a multi-informant measure in that versions of the rating forms have been created for teacher, parent, student, and clinician informants. Since August 2016, the HIDOE has promoted the use of the BIMAS-2 Standard and Flex Teacher-rated forms as their primary measure for data-driven decision-making for SMH services (Matayoshi, 2016). A memorandum from the Office of the Superintendent details that the BIMAS-2 is to be used (a) for baseline and progress monitoring for students who receive counseling as part of an individualized education program (IEP) or 504 plan (i.e., during eligibility process, quarterly

progress monitoring, and at discharge), as well as (b) for school-based behavioral health program evaluation across all HIDOE schools (Matayoshi, 2016). Some benefits of the BIMAS-2 from the perspective of state-level administrators likely include the ease and automation of scoring, interpretation, and progress monitoring through its built-in online platform. Additionally, automated reports can be extracted from this online platform at various levels of data aggregation, including the individual, classroom, grade, and school, making them tailorable for the diverse priorities of various stakeholders. In addition to the BIMAS-2 Standard forms for each informant type, which have standardized items (see Appendix A for the Teacher form), the BIMAS-2 also provides Flex forms, which allow items to be individually created and selected based on the unique concerns of a given student and provide a more tailored approach to progress monitoring. Overall, the HIDOE has been committed to improving school-based mental health services and programs for over three decades, focusing on a number of different efforts (Chorpita & Donkervoet, 2005; Nakasato, 2000). This includes an increased focus on data-based decision-making and progress monitoring of identified students, partnerships with community mental health providers and university researchers, and an investment in evidence-based prevention and early intervention efforts. While the HIDOE does not require schools to implement specific programs or screening protocols, the state has invested in numerous implementation efforts over the last two decades focused specifically on promoting SWPBIS and SEL in schools (Horner et al., 2009; Nakasato, 2000). Although the HIDOE does not currently mandate universal screening or describe it in their overall strategic plan, the BIMAS-2 has been the primary screening, evaluation, and progress measure provided to local schools and was the primary measure promoted by HIDOE leadership for schools pursuing universal social, emotional, and behavioral health screening efforts at the time of data collection for the current

study (i.e., 2016-2018 academic years; A. Bardos, D. Royer, & K. Stern, personal communication, February 17, 2019; Hackel & Kan Hui, 2017).

The Current Investigation

The elementary school included in this study is one HIDOE school that implements a combined SWPBIS and SEL prevention model. As part of that effort, HIDOE psychologists recognized the need to incorporate a systematic universal screening tool (i.e., BIMAS-2) into their multi-source approach for early identification of youth at risk for social, emotional, and/or behavioral challenges. This multi-source identification approach includes ongoing monitoring of student academic outcomes (e.g., grades, achievement scores), ODRs, attendance records, and teacher referral information. During implementation planning for a systematic screening procedure, however, HIDOE psychologists realized that although the BIMAS-2 was designed for use as a universal screener and promoted by HIDOE for progress monitoring of identified students, the 34-item measure was too time consuming and cumbersome for their teachers to sustainably utilize at the schoolwide level. Additionally, it did not adequately meet their objective of a universal screening measure for identifying students at risk of social, emotional, and behavioral concerns, as well as assessing student functioning across CASEL's (2015) SEL competencies and statewide general learning objectives (GLOs) promoted by HIDOE.

When the UBS was initially developed in 2013, there was only one relatively brief universal screening measure available that used rating scales, was teacher report, and aligned with CASEL's (2015) model of social and emotional competencies (Elliott et al., 2018). This measure was the Devereux Student Strengths Assessment- Mini (DESSA-Mini; Naglieri et al., 2011/2014; Nickerson & Fishman, 2009). The original DESSA is a 72-item standardized, norm-referenced, strength-based behavior rating scale measure designed to assess social-emotional

competence in youth grades kindergarten through 8th grade. The DESSA-Mini is an eight-item abbreviated version of the full measure designed to provide a snapshot of students' social-emotional competence for use in universal screening and progress monitoring. There are four forms available for the DESSA-Mini, each composed of unique sets of items which can be rotated to reduce practice effects. The DESSA-Mini has demonstrated good internal consistency (i.e., $\alpha > .90$), alternate form reliability ($r \geq .90$), and test-retest reliability ($r > .88$), across all four forms (Naglieri et al., 2011/2014; Naglieri et al., 2011). As summarized in LeBuffe and colleagues (2018), the DESSA-Mini has also demonstrated evidence of criterion and construct validity, with the technical manual reporting strong correlations between DESSA-Mini Social-Emotional Total (SET) and DESSA Social Emotional Composite (SEC) scores ($r = .95-.96$); adequate accuracy indices for concurrent criterion validity when compared to the 72-item DESSA (i.e., 62-81% sensitivity rates, 83-98% specificity rates, PPVs of 92-97%, NPVs of 86-92%, and area under the curve of 0.79-0.80); and some evidence for predictive validity with one study finding students scoring in the *Need for Instruction* range on the DESSA-Mini in October having a 4.5 times greater likelihood of having a serious disciplinary infraction by end of year (Shapiro et al., 2017). While these psychometric findings of the DESSA-Mini are promising, it is important to note that all analyses except for those concerning predictive validity by Shapiro and colleagues (2017) were obtained using the same sample used to develop DESSA norms, and the scores used to examine the DESSA-Mini and full version came from a single administration of a single form. Thus, more research seemed to be needed to further establish evidence for the reliability and validity of DESSA-Mini scores. In addition, this measure is copyrighted and must be purchased by a school for use.

There is one other relatively *brief*, SEL-focused universal screening measure currently in use that uses teacher rating scales and explicitly maps on to CASEL’s core competencies: Social Skills Improvement System SEL Edition Brief Teacher rating scales (SSIS SEL*b*-T; Anthony et al., 2020); however, this SEL-focused edition of the SSIS was not yet established at the time of data collection for the current study. This is a 20-item version of the larger 51-item SSIS SEL RF-T (i.e., Rating Forms – Teacher), and has been found to function very similarly to the parent measure, with results of an initial psychometric examination suggesting preliminary evidence of reliability and validity of SSIS SEL*b*-T scores across four of five subscales (Anthony et al., 2020). Like the DESSA-Mini, more research is needed to establish psychometric evidence of this brief measure. Furthermore, while it is relatively brief compared to other screening measures, 20 items may still be considered too burdensome by some teachers and the overall cost of the measure is fairly high for schools, especially if purchased with the larger SSIS package.

Due to these feasibility issues and the lack of affordable, well-validated SEL-focused universal screening measures, psychologists were prompted to turn to other creative strategies to implement universal screening that would be free and easy to use, such as their own locally developed nine-item measure, the UBS. The UBS was designed to align with the statewide general learning objectives (GLOs) and core SEL competencies (i.e., CASEL, 2015; Zins et al., 2007), and has been used as a student progress measure as well as a program evaluation assessment of schoolwide GLO and SEL efforts. GLOs are overarching life skills the HIDOE has identified as important to holistic student development and success. These also overlap to some degree with the common SEL competencies described by CASEL (2015) and Denham (2005) and include being a (a) self-directed learner, (b) community contributor, (c) complex thinker, (d) quality producer, (e) effective communicator, and (f) effective and ethical user of

technology (HIDOE & BOE, 2016). While there are a couple brief screening measures available to universally assess student functioning across CASEL (2015) SEL competencies (e.g., DESSA-Mini, Naglieri et al., 2011/2014; SSIS SELb, Anthony et al., 2020), there are no psychometrically evaluated measures designed to assess student performance across GLOs important to educational initiatives within the state of Hawai‘i. As such, the UBS provides a more feasible option for free and brief SEL-based universal screening, and also has the potential to fill an important gap in statewide assessment needs (i.e., formalized assessment of GLOs). Because of the aspirational fit of this measure to the state’s and school’s priorities, its high feasibility as a brief and easily completed measure, and its ease of implementation into already established procedures (i.e., repeated schoolwide UBS administration already implemented for other HIDOE accountability purposes), there was significant stakeholder buy-in for using the UBS for universal screening. Most importantly, the UBS was seen as a good fit for the school’s screening process objectives, which were to identify students who needed more than universal prevention efforts as evidenced by deficits in GLOs and social and emotional skills. Indeed, prior to this study, UBS developers received positive feedback from important school stakeholders to confirm the perceived usability and appropriateness of the UBS; however, the measure was not yet validated in terms of its technical adequacy.

The current study represents an initial psychometric investigation of this locally developed brief screening tool for social, emotional, and/or behavioral challenges in elementary school youth. Given the school’s feasibility concerns with using the BIMAS-2, which was purchased and promoted by the state for universal screening and monitoring, one practical question I aimed to begin answering was the extent to which the UBS might be used in place of the BIMAS-2 for universal screening efforts. Said another way, one goal of the current study was

to understand the extent to which the UBS identifies students at risk of social, emotional, and/or behavioral concerns in a similar manner as the BIMAS-2. Overall, this psychometric investigation had four primary aims: (a) examine UBS factor structure via exploratory factor analysis (EFA), (b) investigate UBS total scale and to-be-determined subscale reliabilities (i.e., internal consistency and test-retest reliability), (c) explore the UBS concurrent criterion validity with other potentially related constructs (i.e., risk status based on BIMAS-2 subscale scores), and (d) preliminarily examine UBS short-term predictive criterion validity over a 7-month time interval with BIMAS-2 subscale risk status as the criterion.

The BIMAS-2 was chosen as the criterion measure for this initial psychometric examination for multiple reasons. The first reason relates to feasibility and burden. Since the BIMAS-2 was already used and approved by HIDOE schools, there was no additional burden on school staff to approve and/or be trained in measure administration. This fit better with policy already in place and allowed for a less invasive and disruptive approach to partnering with the school for this study. Second, it was beneficial from a practical standpoint to compare the novel UBS to the measure most likely to be accessible and utilized by this and other HIDOE schools for universal screening efforts. Third, although not yet examined in formal peer-reviewed studies, the BIMAS-2 developers have conducted several evaluations of the reliability and validity of BIMAS-2 scores, suggesting some preliminary evidence for its psychometric properties (McDougal et al., 2011). While it was extremely preferable to use a more extensively validated screening measure as the reference standard for criterion validity examinations, this was balanced against the practical priorities of relevance to the school and feasibility of study implementation. Further, this study was conceptualized as just the first of many potential

investigations needed to establish support for the utility of the UBS in identifying students at risk of social, emotional, and/or behavioral concerns.

Given the untested nature and factor structure of the UBS, there were limited a priori hypotheses for the four aims of this study. However, I will describe a few tentative speculations across study aims based on my knowledge of the items included on the UBS and findings from examinations of similar mental health screening measures. Based on the EFA results in Aim 1, additional hypotheses were proposed for each subsequent aim and detailed in the corresponding results section. The first aim of this study was to explore the factor structure of the UBS through an exploratory factor analysis (EFA). According to UBS developers, UBS items were designed to align with SEL competencies and HIDEOE GLOs. While a couple HIDEOE GLOs appear to overlap with CASEL (2015) SEL competencies (e.g., self-management GLO and CASEL's self-directed learner competency, Effective communicator GLO and CASEL's social problem solving and relationship skills competencies), there is not a one-to-one correspondence between these competencies. Furthermore, looking at specific item content on the UBS, it appears almost half of the items relate to social and emotional skills while the others relate to academic performance and learning-related behaviors. As such, I suspected a two-factor structure might emerge for the UBS along the lines of either social- versus academic-related behaviors and/or CASEL's (2015) SEL competencies versus unique HIDEOE GLOs. Additionally, given the strong associations between social and emotional skills and academic outcomes for youth in the literature (Corcoran et al., 2018; Durlak et al., 2011; SAMHSA, 2002), I anticipated some potential challenges with overall clarity of the factor structure (e.g., cross-loaded items).

The second aim of this study involved exploring two aspects of reliability of the UBS. This included an examination of the whole scale and to-be-determined factor reliabilities (e.g.,

internal consistency) and temporal stability of UBS scores across all four time points (i.e., first, second, third, and fourth quarters of the academic year). I anticipated the temporal stability of the UBS would likely be impacted by two different factors: (a) unknown effects (e.g., maturation, instruction, SEL interventions such as the *MindUP* curriculum) that were ongoing during the time of administration and (b) time interval length between administrations. However, guidelines still recommend examining test-retest reliability in these circumstances, with the understanding that test-retest coefficients conducted in this examination may be lower than other reliability estimates due to these aforementioned factors (AERA et al., 1999; Glover & Albers, 2007).

Many published studies of commonly recommended screeners have also utilized longer test-retest time intervals (i.e., 1.5 months to 3 years). For example, Lane, Menzies, and colleagues (2012) used test-retest intervals of 2 to 8 months in their validation study of the Student Risk Screening Scale for Internalizing and Externalizing (SRSS-IE). Notably, the schools included in their study sample were also implementing a multi-tiered prevention program (i.e., Ci3T) during SRSS-IE data collection similar to the current, proposed study (i.e., PBIS as behavioral component and SEL as social/emotional component within larger tiered prevention approach; Lane, Menzies, et al., 2012). Evaluations of the Behavioral and Emotional Rating Scale – 2 utilized test-retest intervals of 6 weeks to 6 months (Epstein & Sharma, 1998; Epstein et al., 2004), while an examination of the Behavior Assessment System for Children, third edition, Behavioral-Emotional Screening System (BASC-3 BESS; Kamphaus & Reynolds, 2015) Teacher rating scale used test-retest intervals between 7 months and 3 years (Dever et al., 2018). Lastly, Muris and colleagues (2003) utilized a time interval of 2 months for their test-retest reliability analyses of the Strengths and Difficulties Questionnaire (Goodman, 1997). Across these studies, test-retest correlation coefficients generally became weaker as more time passed

(e.g., $r > .80$ for BESS 7-month interval compared to $r = .50$ to $.64$ for intervals spanning multiple academic years; Dever et al., 2018). Interestingly, all test-retest coefficients across these studies were in the large range (Cohen, 1988) regardless of time interval. Although no specific hypotheses were offered for the current study, I suspected the effect sizes of the test-retest reliability coefficients would be weaker for longer test-retest intervals, but still evidence large effect sizes across all intervals, in line with previous studies of other screeners (e.g., Epstein & Sharma, 1998; Epstein et al., 2004; Lane, Menzies, et al., 2012).

The third aim of this study was to examine UBS criterion-related validity patterns with another measure routinely used within HIDOE (i.e., the BIMAS-2). Specifically, convergent validity was explored in terms of the correlation between UBS total and to-be-determined subscale scores and BIMAS-2 subscale scores. Hypotheses guiding the examination of convergent validity patterns were based on patterns identified in the forthcoming UBS factor subscales. For example, if the UBS is found to have a factor characterized by items concerning academic performance, then it would be hypothesized to correlate significantly with the BIMAS-2 Academic Functioning scale. Criterion-related validity patterns will also be examined in Aim 4 related to the power of UBS scores towards the beginning of the academic year to predict risk categorization across BIMAS-2 subscales towards the end of the academic year, which will be explored through a series of logistic regression analyses. Although it is unclear the extent to which UBS to-be-determined subscales will relate to the pre-established BIMAS-2 subscales, both measures are designed as screeners of social-emotional and behavioral disturbance in youth and thus, should be expected to relate in terms of student risk identification.

Method

Participants

Teachers

This project involved archival data from teacher reports of student behaviors across two different screening measures during the 2017-2018 academic year. Data included measures completed by nine general education teachers at a public elementary school in the state of Hawai‘i. The distribution of these nine teachers by grade level is presented in Table 1. According to UBS administrators, there was a 100% participation rate for general education teachers since the universal screening effort was a schoolwide initiative (S. Myers, personal communication, July 31, 2019). These teachers completed the UBS quarterly and the BIMAS-2 biannually (i.e., Quarters 2 [Q2] and 4 [Q4]) on all students in their classroom (ranging from 18-31 students per classroom) during the 2017-2018 academic year. The 2017-2018 academic year was chosen because it was the first year the BIMAS-2 was collected from teachers alongside the UBS, thus providing a way to evaluate the UBS against an established, psychometrically tested screening tool. There was no demographic data available for the teachers included in the study.

Table 1

Distribution of Teachers (N = 9) and Students (N = 230) by Grade

Grade	Teachers <i>n</i>	Students <i>n</i> (%)
Kindergarten	2	43 (18.7)
First grade	2	42 (18.3)
Second grade	1	26 (11.3)
Third grade	1	25 (10.9)
Fourth grade	1	32 (13.9)
Fifth grade	1	33 (14.3)
Sixth grade	1	29 (12.6)

Students

Archival data yielded teacher-rated measures for 230 students (41.3% female¹) ranging from kindergarten through 6th grade (see Table 1 for distribution of students by grade level). Data from all students in general education classrooms at this public elementary school were included in this study. Due to the archival and de-identified nature of the data, only demographic information recorded at the time of data collection were available for the investigation, including gender, grade, and unique teacher ID. I requested additional demographic data from Hawai‘i Department of Education (HIDOE) personnel specifically for my 230 participants, however, these requests were denied due to staff time constraints and challenges connecting data sources post hoc. Thus, I referenced publicly available reports on the elementary school sampled for this study to supplement the limited demographic information available in my dataset. These included enrollment and performance reports produced by the school and the state for the 2017-2018 academic year (see Table 2). Overall, demographic information for the students in this sample were similar to statewide data on all students enrolled in HIDOE in 2017 across all demographic variables in Table 2 except for English learner students (i.e., 20.5% statewide; Civil Rights Data Collection [CRDC], n.d.). The Strive Hawai‘i report (HIDOE, 2018) also included the breakdown of students receiving services through the school across the SWPBIS three tiers of support. The percentages of students receiving supports associated with each tier are included in Table 2 for Quarters 1 and 4. Students’ need for services across these tiered supports were determined by their scores on the to-be-evaluated UBS in combination with staff and parent referrals for supports and other objective metrics (e.g., ODRs, academic performance).

¹ Gender reported throughout this manuscript refers to assigned sex at birth and may or may not represent the gender with which the student identifies. Data related to self-identified gender was not available for the current study.

Table 2*Student Demographic Information*

Demographics	<i>n</i>	%
Gender ^a		
Female	95	41.3
Male	135	58.7
Ethnicity ^b		
Asian	35	15.3
Black	0	0.0
Bi- or Multi-racial	46	20.1
Hispanic	48	20.1
Pacific Islander	65	28.4
White	35	15.3
Eligible for free- and reduced-price lunch ^c	128	59.0
Received special education services ^c	39	18.0
English language learners ^c	7	3.0
Students receiving tiered supports at Quarter 1 ^c		
Tier 1	186	85.6
Tier 2	14	6.5
Tier 3	17	7.9
Students receiving tiered supports at Quarter 4 ^c		
Tier 1	203	93.5
Tier 2	5	2.3
Tier 3	9	4.1

^a Data obtained from school records for the study sample included in this archival dataset ($N = 230$). ^b Data obtained from public enrollment data reported by HIDOE (School Digger, n.d.) and includes $N = 229$ students surveyed from the elementary school in 2017. ^c Data obtained from the 2017-2018 Strive Hawai'i Report (HIDOE, 2018) and includes $N = 217$ students enrolled in the elementary school during the 2017-2018 academic year.

Power

Concerning Aim 1, there are a wide range of recommendations for optimal sample size in factor analysis, with some focusing on blanket sample size requirements regardless of questionnaire size (e.g., 300 participants minimum; Comrey & Lee, 1992); others focusing on observation to item ratios (e.g., 5:1 ratio in samples of at least 100; Streiner, 1994); and others challenging use of across-the-board recommendations. Regarding the last point, Guadagnoli and Velicer (1988) suggest using a combination of the total sample size, the number of items per factor, and the magnitude of factor loadings to determine adequate sample size. In a larger study of factor analysis procedures, Guadagnoli and Velicer (1988) found that larger sample sizes (i.e., the 300-400 blanket recommendation) were only needed if the factor loadings were as low as .40, while smaller sample sizes (i.e., $100 < N < 150$) were found to be acceptable if data indicated factor loading scores higher than .80, and slightly larger sample sizes (i.e., $N > 150$) were needed if factor loadings were in the .60 range. Based on these guidelines, the recommended sample size for this study could range from a minimum of 100 to up to 300 participants depending on the size of factor loadings, with a moderate sample size of 150 sufficient if the factor loadings for UBS items met the .60 cut off (Guadagnoli & Velicer, 1988).

With regards to the correlational analyses described in Aims 2, 3, and 4, G*power (Faul et al., 2009) was used to estimate sample sizes set for a two-tailed test at an alpha of .05, a power level of .80, with effect sizes set at small, medium, and large (Cohen, 1988). For the bivariate correlations proposed in Aims 2 and 3, G*power analyses indicated a sample size of 29, 84, and 782 would be needed for correlation analyses based on an effect size of 0.5 (i.e., large), 0.3 (medium), and 0.1 (i.e., small), respectively.

Lastly, regarding the logistic regression analyses in Aims 3 and 4, sample size estimations were based on commonly cited recommendations in the literature. This method was chosen due to the unknown nature of some parameter values needed to run the G*power analysis (e.g., odds ratio, which was unavailable due to the novel nature of the measure and unknown subscale constructs). Peduzzi and colleagues' (1996) commonly cited equation for sample size estimation with logistic regression recommends a sample size based on the equation $N = 10k/p$ where k = the number of predictors/covariates and p = the smallest of the proportions of not at-risk or at-risk cases in the population, with a minimum acceptable sample size of 100. In terms of the number of predictors/covariates (i.e., k), this was set at one (i.e., UBS continuous score). In terms of the smallest of proportions, it is estimated that approximately 15-20% of youth will likely be at-risk in a given school system based on the typical estimation that approximately 80-85% of students can be managed through only Tier 1 supports (e.g., Gresham et al., 1998). Based on this, p for the current study could be estimated at a low of 0.15. Utilizing the equation from Peduzzi et al. (1996), this would indicate a recommended sample of 66 (i.e., $N = 10*1/0.15$) for a single predictor. To examine potential covariates, such as gender, a sample size of 134 (i.e., $N = 10*2/0.15$) was required. Given the minimum acceptable sample size was recommended at 100 (Peduzzi et al., 1996), the requisite sample size range for my proposed logistic regression analyses was around 100 students for simple logistic regression analyses and 134 students to include gender in the model. In summary, estimates across all aims generally recommended a sample size of 100 to 300, depending on the conditions of the data. Considerably more participants were needed for correlational analyses if there were small effect sizes (i.e., $N = 782$).

Measures

Behavior Intervention Monitoring Assessment System 2, Teacher Standard Form (BIMAS-2)

The BIMAS-2² Standard Form (McDougal et al., 2011; see Appendix A) is a 34-item multi-informant (i.e., teacher, parent, student, clinician) measure that was designed for use in evaluation and universal screening of youth aged 5-18 across behavioral, emotional, social, and academic domains. The format of the forms is identical across all informant versions except for the Clinician Standard form. The BIMAS-2 Teacher Standard Form asks teachers to rate how often a given behavior occurred for the student in the past week using a 5-point scale ranging from 0 (*Never*) to 4 (*Very Often*). For ratings indicating the presence of a behavior, descriptions are provided to clarify the anchors, including: 1 (*Rarely*) = observed 1-2 times or to a minimum extent, 2 (*Sometimes*) = observed 3-4 times or to a moderate extent, 3 (*Often*) = observed 5-6 times or to a significant extent, and 4 (*Very Often*) = observed 7 or more times or to an extreme extent. The BIMAS-2 consists of five separate subscales (see Appendix B), including three problem-oriented Behavioral Concern scales: (a) Conduct (i.e., 9 items relating to externalizing symptoms), (b) Negative Affect (i.e., 7 items relating to internalizing symptoms), and (c) Cognitive/Attention (i.e., 7 items relating to attention, impulsivity, and executive functioning), and two strength-based Adaptive scales: (d) Social (i.e., 6 items relating to interpersonal skills and communication) and (e) Academic Functioning (i.e., 5 items relating to academic performance and attendance). There is no total score for the BIMAS-2.

² The authors of the BIMAS-2 do not specify the difference, if any, between the BIMAS original and BIMAS-2 in any of their published works on the BIMAS. The BIMAS developers have no known technical manual specifically developed for the BIMAS-2 and continue to reference the original BIMAS Technical Manual (McDougal et al., 2011) in all BIMAS-2 correspondence. Thus, it is assumed that there are no differences between the two BIMAS versions in item content or scoring.

Informants complete the BIMAS-2 using the electronic administration capabilities of the BIMAS-2 Online data management system (BIMAS-2 Online). All scoring and interpretation are conducted using the web-based platform, which generates different types of reports from the scored data at both the individual- and multi-student aggregate levels. The multi-student reports assess and monitor student behaviors at the level of group, class, grade, school, and/or district. Four different reports can be generated for the BIMAS-2, including an assessment report, which will be used in the current study and displays the results from a single time point of administration at either the multi-student level or individual level. Descriptions of the other available reports can be found in the BIMAS Technical Manual (McDougal et al., 2011). For the current study, data for each time point will be taken from individual level assessment reports for BIMAS-2 Teacher Standard forms.

The interpretations of BIMAS-2 results are provided via *T*-scores across scales based on nationally derived norms (McDougal et al., 2011). For the Behavioral Concern Scales (i.e., Conduct, Negative Affect, and Cognitive/Attention), higher *T*-scores correspond with greater levels of concern: *T*-scores above 70 are interpreted as *high risk* and indicate a much higher number of concerns than is typical, *T*-scores between 60 and 69 are interpreted as *some risk* and should be further reviewed to determine if further assessment or intervention is warranted, and *T*-scores below 60 are interpreted as *low risk* and likely not warranting further action. For the Adaptive Scales (Social, Academic Functioning), lower *T*-scores correspond with greater levels of concern: *T*-scores above 60 for a given scale indicate *a strength*, *T*-scores between 41 and 59 are interpreted as *typical*, and *T*-scores 40 and below indicate *a concern* and further assessment and/or intervention may be needed.

Most research regarding the psychometric properties of the BIMAS and/or BIMAS-2 are found in the technical manual for the BIMAS published by McDougal and colleagues (2011). In my independent review of the literature using Google Scholar, I found a total of 30 papers that cited the original technical manual. Of these 30 citations, none were peer-reviewed publications that involved a psychometric investigation of the measure. Two unpublished doctoral dissertations reported examining the validity of the BIMAS-2 (Marandos, 2020; van Luling, 2015); however, both studies compared BIMAS-2 performance against teacher nomination and not against another valid and reliable identification method. Thus, the results of these studies still do not provide an adequate external replication of the validity of the BIMAS-2 in identifying students at risk of social, emotional, behavioral, or academic concerns. Most other studies found in my literature review either included the BIMAS-2 as one of the study measures but did not report on any psychometric statistics (e.g., Cronbach's alpha) or included only a description of the BIMAS-2 within a larger review of screening measures.

The following discussion of psychometric characteristics of the BIMAS-2 focuses on the investigations summarized in the BIMAS Technical Manual (McDougal et al., 2011) that were conducted by measure developers. Evaluations of the BIMAS-2 Standard version (c.f. BIMAS-2 Flex) utilizing Teacher, Parent, and Self-report forms have suggested adequate reliability and validity across its subscales. Examinations of internal consistency reliabilities for the BIMAS-2 Teacher report form suggest Cronbach's alpha coefficients in the excellent range for Conduct ($\alpha = .91$) and Cognitive/Attention ($\alpha = .91$) subscales and in the good range for the Negative Affect ($\alpha = .85$), Social ($\alpha = .85$), and Academic Functioning ($\alpha = .81$) subscales using a weighted sample of 85% normative and 15% clinical cases, mixed for real-world relevance (McDougal et al., 2011). It appears that all analyses of BIMAS subscale reliability and validity were conducted

with the same normative and clinical samples. Normative samples were described as “representative of the general US population in terms of age, gender, race, geographic location, and parental education level in accordance with the 2000 US Census” (McDougal et al., 2011, p. 66). According to McDougal and colleagues (2011), the clinical sample included youth who had received a primary clinical diagnosis by a qualified mental health professional through appropriate assessment of formal criteria for the disorder. Evidence is provided in the manual suggesting adequate standard error of measurement (*SEM* values = 2.58 to 4.36 across subscales), test-retest reliability over a 2- to 4-week interval ($r = .85$ to $.91$, $N = 112$ teacher ratings), and inter-rater reliability ($r = .54$ to $.69$; $N = 162$ teacher-student pairs; $r = .79$ to $.86$, $N = 162$ teacher-parent pairs) for *T*-scores across BIMAS-2 teacher subscales.

Regarding validity of the BIMAS-2 Teacher form, the manual presents findings related to content, convergent, and divergent validity. The authors make an argument for the content validity of the BIMAS-2 by citing studies spanning from 1994 to 2004 (Meier, 1997, 1998, 2000, & 2004). These studies examined the content validity of the BIMAS-2 by exploring the relations of individual items to external criteria, conducting content reviews, and examining feedback from colleagues working in the field (McDougal et al., 2011). Empirical support for the BIMAS-2 scale structure was established via a series of confirmatory factor analyses with a combined sample of both normative ($n = 1,400$ teacher ratings) and clinical ($n = 538$ teacher ratings) samples of youth using maximum likelihood, generalized least squares estimation. Results indicated adequate fit of the final BIMAS-2 five subscale model: Normed Fit Index = $.91$, Non-Normed Fit Index = $.87$, Comparative Fit Index = $.91$, and Steiger-Lind Root Mean Square Error of Approximation Index = $.13$ (McDougal et al., 2011).

As described in the manual, convergent and divergent validity has also been examined for the BIMAS-2 Teacher form via comparisons of BIMAS-2 subscales to relevant scales on the Conners Comprehensive Behavior Rating Scales (Conners, 2008). In terms of screening ability, BIMAS developers have conducted several different examinations. First, they compared mean *T*-scores from their clinical sample to their normative sample and found Cohen's *d* values between |1.0| to |1.7| across all subscales, with all comparisons in the expected direction, with clinical sample scores indicating poorer functioning. Also, more in-depth analyses comparing subscale *T*-scores between subsamples of the clinical sample against the normative sample also performed as expected. For instance, the subsample of the clinical sample presenting with disruptive behavior had a larger mean score on the Conduct subscale when compared to normative counterparts (Cohen's *d* = 2.1). Similar patterns were found for the Cognitive/Attention subscale with the clinical subsample with AD/HD (Cohen's *d* = 1.9), the Negative Affect subscale with the clinical subsample diagnosed with either anxiety (Cohen's *d* = 2.1) or depression (Cohen's *d* = 2.2), and the Social subscale with youth presenting with pervasive developmental disorder (Cohen's *d* = -1.8) when compared to the normative sample means for those subscales.

The authors also explored the classification accuracy of the BIMAS-2 Teacher forms *T*-scores to determine clinical/non-clinical group categorization through a series of discriminant function analyses (DFAs). These analyses explored the extent to which BIMAS *T*-scores (using cut-offs for risk and concern status) predicted group membership in the clinical and non-clinical groups. As reported in the manual (McDougal et al., 2011), the authors found the following classification indices: 82.5% overall correct classification (i.e., percentage of correct group classifications of normative or clinical categories), 80.1% sensitivity (i.e., percentage of clinical cases correctly predicted), 83.4% specificity (i.e., percentage of normative cases correctly

predicted), 64.9% positive predictive power (i.e., percentage of students accurately identified as having a clinical condition), and 91.6% negative predictive power (i.e., percentage of students accurately identified as not having a clinical condition; McDougal et al., 2011).

In the current study, the Cronbach's alpha coefficients for the five BIMAS-2 subscales at Q2 were in the "good" to "excellent" range ($\alpha = .79$ to $.93$) and in the "acceptable" to "excellent" range for Q4 data ($\alpha = .65$ to $.81$; see Table 3). Item-level statistics for the BIMAS-2 subscales are reported in Appendix C. Overall, several items had very low means suggesting low base rates with regard to teacher endorsement, particularly for items on the Conduct (i.e., Q2 $M = 0.12$ to 0.55 ; Q4 $M = 0.01$ to 0.46) and Negative Affect (i.e., Q2 $M = 0.14$ to 0.89 ; Q4 $M = .04$ to $.78$) subscales (valid range = 0 to 4). Using the current study sample, test-retest reliability for BIMAS-2 subscale scores over an interval of 5 months, 10 days was found to be adequate, with Pearson correlation coefficients evidencing "large" effect sizes (Cohen, 1988) for all subscales (T -score $r = .50$ to $.79$; $p < .01$).

Table 3*Reliability Coefficients for BIMAS-2 Teacher Form Subscales at Q2 (n = 205) and Q4 (n = 211)*

BIMAS-2 Subscale	Q2 No. of Items	Q2 Cronbach's α	Q4 No. of Items	Q4 Cronbach's α	Q2 – Q4 test-retest r
Conduct	7 ^a	.827	8 ^b	.812	.63*
Negative Affect	7	.836	7	.771	.66*
Cognitive/Attention	7	.931	7	.910	.79*
Social	6	.789	6	.651	.50*
Academic Functioning ^c	5	.794	5	.795	.70*

Note. Table displays Cronbach's alpha coefficients for Q2 and Q4 administrations in the current study, as well as test-retest correlation coefficients for each subscale.

^a Items 29 and 32 excluded from reliability estimate due to zero variance. ^b Item 29 excluded from reliability estimate due to zero variance. ^c Two items on the academic functioning subscale were reverse scored prior to calculating Cronbach's α coefficients.

* $p < .01$.

Universal Behavior Screener (UBS)

The UBS (S. Myers & K. McDonald, personal communication, 2019; see Appendix D) is an unpublished nine-item, teacher-report brief screening measure designed by local psychologists within the HIDOE to assess elementary school students' social-emotional and behavioral well-being, particularly over time in response to Tier 1 interventions. The UBS is a rationally derived screener developed over the course of four academic years (i.e., 2013-2014 to 2016-2017). A rational or theoretical approach to test construction uses "a theory of the target construct to inform the development of items" (Ruscio, 2014, p. 1). The items on the UBS correspond to General Learner Outcomes (GLOs) across HIDOE schools and to common SEL

competencies (see CASEL, 2015) that are also found within three common Social Emotional Learning (SEL) curricula used in the HDOE: *Mind Up* (The Hawn Foundation, 2011), *Second Step* (Committee for Children, 2011), and *Rainbows in Me* (Kawamoto, 2016). The final nine items included on the UBS were determined via a multi-step process involving iterative feedback on measure length, format, and content by members of school-based mental health teams (e.g., school psychologist, clinical psychologist, behavioral health specialist, teachers, counselor) at two elementary schools within the State of Hawai'i Department of Education (HDOE). For example, teacher complaints about the length of the original measure and cumbersome nature of its resultant data led to an iterative reduction of UBS from its original 27 items to the current nine-item measure.

The UBS asks teachers to rate each student in their class on a scale of 1 to 5 based on how they behaved within the past month in comparison with expectations of their age, with each rating corresponding to typical grade-level report card ratings (i.e., Deficient, Well Below, Developing Proficiency, Meets Proficiency, and Meets with Excellence). For example, as seen in Appendix D, a rating of 1 = *Deficient in performing the expectation/the behavior never occurs*, whereas a rating of 5 = *Meets with Excellence in performing the expectation above the typical/average student*. Currently, the UBS is scored by calculating the average of all nine items to create one overall composite score for each student, indicating a possible range of scores of 1.0 to 5.0, with higher ratings indicating greater degree of functioning and lower ratings indicating deficits in functioning.

Due to the untested nature of this novel measure, there are currently no available norms or associated cut-offs for scoring and interpretation. For the 2017-2018 breakdown of students across tiers (see Participants section), measure developers based risk status classifications on

each student's deviation from their classroom mean (i.e., *Priority* if UBS score = 1.5+ standard deviations from classroom mean, *Near Priority* if UBS score = 1.0-1.49 standard deviations from classroom mean). This scoring and interpretation method was designed to help account for potential teacher rating bias and allow for identification of students relative to their classroom peers. Changes in students' scores between repeated administrations were used as an additional form of identification throughout the school year such that students who had a loss in their UBS score of at least 0.5 standard deviations at any subsequent administration were flagged for follow up and closer evaluation for needs and supports. Thus, the repeated administration of the UBS allowed for multiple opportunities for a student to be identified throughout the school year.

During the 2017-2018 academic year, UBS developers completed routine annual teacher administration and youth identification procedures as outlined above (S. Myers. & K. McDonald, personal communication, March 2019). During this initial UBS administration, 85% of students' scores fell at or less than 1.0 standard deviations from the class mean, indicating no need for additional follow up. The other approximately 15% of students were split between *Near Priority* and *Priority* status (i.e., 6.51% and 7.91%, respectively) based on the aforementioned standard deviation cut-offs.

To date, no psychometric evaluation has been conducted on the UBS. For the current study, UBS total scale and to-be-determined subscale scores were calculated as the average of scores across items included on that subscale, ranging from 1.0. to 5.0. These average scores were used in all analyses of the UBS. The UBS developers' standard deviation cut-off method for identifying student risk status was not used in the current study.

Procedures

Clinical and school psychologists from the HDOE coordinated teacher administration of the UBS at a HDOE elementary school from 2016-2020 as part of their routine educational and administrative practices. Beginning in August 2016 (academic year 2016-2017), all participating teachers completed the UBS on all students in their classroom each quarter using an online Google spreadsheet they were emailed by the UBS developers at each time of administration (S. Myers, personal communication, September 27, 2019). This Google spreadsheet was formatted exactly like the paper version included in Appendix D, with identical instructions and listed all students' names in alphabetical order with any newly enrolled students added at the end of the list for each quarter. Once completed, the teachers sent the spreadsheet to the UBS developers to be aggregated into the schoolwide Excel spreadsheet database. Since August 2017, teachers also completed the BIMAS-2 on all students in their classroom twice annually (i.e., Q2 and Q4) using the BIMAS-2 online data management system. The approximate dates of administration for each quarter were as follows: Quarter 1 (Q1): October 4, 2017, Quarter 2 (Q2): December 6, 2017, Quarter 3 (Q3): March 7, 2018, and Quarter 4 (Q4): May 16, 2018 (S. Myers, personal communication, August 5, 2019). During this time, the elementary school was employing *MindUP* (The Hawn Foundation, 2011) at the school-wide level, which is a SEL curriculum designed to include four units of 15 lessons throughout the schoolyear. The implementation of this curriculum, however, was not formally tracked by teachers or administration and thus the extent to which this intervention was received by all students in the current study is unknown. All UBS raw data and BIMAS-2 Teacher Standard Form raw data and *T*-scores were compiled in an excel database starting with the first administration in 2016, hand-entered and checked by HDOE psychologists (S. Myers, personal communication, March 2019). Along with UBS and

BIMAS-2 data, this database included other pertinent information about administration (e.g., administration time point, gender and grade level of the youth, and unique IDs for both teachers and students). After approval by the HIDOE Data Governance Office, I received a copy of the de-identified archival dataset from the UBS developers, which included data from the 2016-2017 and 2017-2018 academic years. The policies and procedures of the current study have been approved by the University of Hawai'i at Mānoa Committee on Human Studies (CHS #2019-00009, 3/21/2019) and a data sharing agreement with HIDOE was approved for all study activities by the HIDOE Data Governance Office on July 26, 2019.

Analytic Strategy

Data Diagnostics and Preparation

All data were initially inspected using IBM's Statistical Package for the Social Sciences (SPSS) Version 23 for Windows (2015). All data used in this study were examined for impossible values across all items, subscales, and total scores of measures. Potential data entry errors were resolved through consultation with UBS developers who maintained access to the fully identified dataset.

The scope of missing data was expected to be low given measure completion was a part of routine educational practices expected of teachers by school administration. Furthermore, all measures were completed under the oversight of district-level clinical and school psychologists who required complete data for student monitoring purposes. Psychologists at the participating school confirmed teachers' response rates in terms of completing both the UBS and BIMAS-2 for all enrolled students at the time of administration for each quarter. Due to explainable changes in students' enrollment statuses throughout the year (e.g., students enrolling after start of school year, students moving away or transferring schools), missing data were only considered

non-responses (i.e., truly “missing”) if the student had data for some but not all items at a given time point (e.g., only eight of the nine UBS items were completed at Q1). Students missing nine of nine items at a given time point were assumed to not be enrolled at that time, and thus, data were not considered missing. UBS and BIMAS-2 missingness were assessed by calculating frequencies at all levels of analysis (i.e., item-, subscale-, and/or total scale-levels, within and across measures). Management of missing data was based on recommendations in the literature for the size and scope of missingness, as well as the impact on relevant analyses.

Standard distributional properties of the data were examined at the item-level for Aim 1 (e.g., skew, kurtosis, Shapiro-Wilk’s *W* test of normality). Following the results of the EFA on the UBS factor structure, additional examinations of the data were conducted for analyses in Aims 2 through 4. First, standard distributional properties of the data were examined at the total scale- and/or subscale-level for the UBS and BIMAS-2 at each time point, including skewness, kurtosis, extreme values, and statistical tests of normality. Examination of normality was relevant for all analyses across aims except for the logistic regression analyses in Aims 3 and 4 (Field, 2018). Skewness and kurtosis were regarded as “excellent” and “acceptable” if the statistic was between -1.0 to 1.0 and -2.0 and 2.0, respectively (George & Mallery, 2003). Shapiro-Wilk’s *W* statistic (Shapiro & Wilk, 1965) was used to statistically test data normality with $p < .05$ suggesting a significant deviation from normal distribution. Sample size was considered when interpreting the results of normality tests for a couple reasons. First, the central limit theorem posits normality can be assumed with large samples. According to Field (2018), as long as there is not a large amount of skew or kurtosis, the threshold for a “large” sample may range from 30 to 160 participants. Second, Shapiro-Wilk’s *W* statistic is influenced by sample

size such that it may be significant for even small, inconsequential effects when the sample is large (Field, 2018).

Potential outliers were formally identified at the total scale- and/or subscale-level through examination of box plots and z -scores using IBM SPSS Version 23 (2015). Absolute values of z -scores were examined based on benchmarks expected in a normal distribution: extreme cases ($|z| > 3.29$; very few cases expected), probable outliers ($|z| > 2.58$; $\leq 1\%$ cases expected), potential outliers ($|z| > 1.96$; $\leq 5\%$ cases expected), and cases in the normal range ($|z| < 1.96$; approximately 95% cases expected). The current sample included all students enrolled at the school; thus, it was anticipated that students' scores indicating greater concerns (i.e., lower scores on UBS total scale and to-be-determined subscales, lower T -scores on BIMAS-2 Adaptive scales, and higher T -scores on BIMAS-2 Behavioral Concern scales) might cause high skewness and kurtosis values, or be identified as outliers due to lower base rates of mental, social, and emotional challenges from youth in a normative sample (Costello et al., 2003). Given these cases represented important information for the current study, outliers were retained if they did not represent data errors and transformations were not utilized for the data (Tabachnick & Fidell, 2007). In the event of substantial outliers or issues with non-normality in the data, other statistical methods robust to non-normality were considered as alternatives to proposed data analyses.

Prior to conducting any analyses related to Aims 2-4 of the current study, descriptive statistics were examined on the UBS and BIMAS-2 measures, including minimums, maximums, means, and standard deviations of relevant subscale and total scale scores across all four quarters.

Inclusion Criteria for Study Aims. Inclusion criteria were specific to each of the analyses across the four study aims to maximize the number of records used for data analyses and ensure appropriate comparisons were made. These priorities are explained for each of the analyses described below. For all four study aims, students were only included in analyses if they had 100% subscale and/or total scale data available for the relevant measure(s) and time point(s) utilized.

For the first study aim (i.e., EFA), only students with UBS data completed at Quarter 2 (Q2) were included. This time point was chosen for the sake of consistency, as Q2 was included in the majority of study aims. Item-level analyses were explored to confirm that Q2 data were similar to the other three time points and thus, did not represent an outlier, as well as identify any potential issues unique to Q2 data that are not representative of typical UBS performance and thus, suboptimal for use in the EFA (e.g., unique problems with inconsistent means across items, low item-total correlations, excessive skewness and kurtosis, infrequently used items; Lane, Oakes, et al., 2012).

For Aim 2 (i.e., reliability analyses) internal consistency analyses, the calculation of Cronbach's alpha coefficients was done separately for each quarter. Thus, inclusion criteria were separate for each time point, with students included in each reliability calculation if they had UBS data for that quarter. For temporal stability (i.e., zero-order bivariate correlations), pairwise deletion strategies were utilized for analyses. For example, for the correlation between Q1 and Q2, only students who had data available for both Q1 and Q2 were included in the analysis. For Aim 3 (i.e., criterion-related validity), students were included in all analyses if they had data for both the UBS and BIMAS-2 at Q2. This time point was chosen because it is the first quarter for

which BIMAS-2 data were collected. For Aim 4 (i.e., predictive validity), students were included in analyses if they had data available for both the UBS and BIMAS-2 at both Q1 and Q4.

Calculating BIMAS-2 Risk Status Categories for Aims 3 and 4. Prior to conducting logistic regression analyses for Aims 3 and 4, BIMAS-2 binary risk status variables were created for BIMAS-2 subscales at Q2 and Q4. As previously described, interpretation of BIMAS-2 *T*-scores correspond to three levels risk or concern for all subscales. For the logistic regressions in the current study, two of the three categories were combined to create binary risk status for each subscale. For BIMAS-2 Behavioral Concern Scales (i.e., Conduct, Negative Affect, and Cognitive/Attention), risk status was broken into “any risk” and “low risk” categories. *Any risk* corresponded to BIMAS-2 *T*-scores ≥ 60 , combining BIMAS-2’s *high risk* (i.e., *T*-scores > 70) and *some risk* (i.e., *T*-scores between 60-69) categories. For the current study, *Low risk* corresponded to the BIMAS-2 *low risk* category (i.e., *T*-scores < 60). This was done to create more equal groups for comparisons and ensure the examination of the UBS prioritized sensitivity to any level of risk rather than just one level of risk, given its purpose as an initial gate of identification. Regarding the balance of groups for comparisons, McDougal and colleagues (2011) found that average *T*-scores for the clinical population in their study fell in the *some risk* range across Behavioral Concern Scales, with a lower frequency of students scoring in the *high risk* range. Given the already lower base rates of students needing additional supports compared to those who do not, it was beneficial to combine risk groups to bolster group size. This mirrors methodology and rationale used in psychometric investigations of other screening measures with multiple risk classifications, such as the SRSS-IE (Lane et al., 2019) and the BESS-Teacher (Dever et al., 2018). For BIMAS-2 Adaptive Scales (i.e., Social, Academic Functioning), risk status was broken into “concern” and “no concern” categories. *Concern* in the current study

corresponded to the BIMAS-2 *concern* category (i.e., T -scores ≤ 40), while *no concern* was defined as BIMAS-2 T -scores > 40 , combining BIMAS-2's *strength* (i.e., T -score > 60) and *typical* (i.e., T -score between 41-59) categories. *Strength* and *typical* categories were combined since they were believed to represent the absence of concern for Adaptive subscales, whereas *concern* reflects a potential need for intervention. The approach described here for both the BIMAS-2 Behavioral Concern and Adaptive Scales was also used by McDougal and colleagues (2011) in their initial validation studies of the BIMAS-2.

Examining the Influence of Gender on BIMAS-2 Scores. Hypothesized covariates were examined prior to conducting the logistic regression analyses for Aims 3 and 4. For the current study, the impact of student gender on BIMAS-2 scores were explored using t -tests for both Q2 and Q4 administrations. In the event significant differences in scores were found between groups, modifications to proposed analyses were considered, such as multiple logistic regression.

Proposed Analyses

All analyses across Aims 1 through 4 were conducted using IBM SPSS Version 23 for Windows (2015) except the logistic regression analyses in Aims 3 and 4, for which SPSS Version 27 for Macintosh (2020) was used.

Aim 1: Examine UBS Factor Structure. Toward this first aim, an exploratory factor analysis (EFA) was conducted using all nine items of the UBS at Q2 to determine the factor structure of the measure. Decisions about item retention were determined based on both statistical and theoretical considerations. First, the appropriateness of conducting an EFA with the data were assessed through examinations of assumptions, factorability, and sampling adequacy. Parallel analysis with raw data permutations was utilized to preliminarily determine

cut points for the number of factors prior to running the common factor analysis. This analysis involved comparing each eigenvalue against an eigenvalue for the corresponding factor in 1,000 parallel datasets generated from the raw data (O'Connor, 2000). This method has been deemed preferable to both Kaiser's criterion, which is known for risking overestimation of factors, as well as the basic scree plot of the raw data (Zwick & Velicer, 1986). Following the parallel analysis, a common factor analysis was conducted using the principal axis factoring method. Factors were extracted based on multiple considerations, including the results of the parallel analysis, scree plot evaluation and percent of variance explained by each factor, and Kaiser's criterion of eigenvalues greater than 1.0 (Fabrigar et al., 1999; Field, 2018). Once factors were extracted, items were subjected to oblique rotation, which is generally preferable to orthogonal rotation in social sciences research due to the likelihood of correlation among factors related to behavioral constructs (Costello & Osborne, 2005). Factor loadings were also examined to confirm that all items significantly loaded onto their respective factors. Items with high loadings ($> .35$) were retained (Floyd & Widaman, 1995) and considered for deletion if they loaded onto more than one factor (Costello & Osborne, 2005). In addition to these statistical procedures, decisions about item retention were also based on theoretical considerations. For example, examining the extent to which the retained items within factors shared conceptual meaning, the extent to which retained items between factors suggested different and distinct constructs, and whether there were at least three items on each factor with significant factor loadings. This proposed methodology mirrors that used in similar evaluations of screening measures (e.g., Lane, Menzies, et al., 2012; Lane et al., 2016). The factor structure identified in the EFA was used in Aims 2-4. Following identification of any subscales on the UBS, exploratory hypotheses were

proposed for analyses conducted in Aims 2, 3, and 4 (i.e., reliability, concurrent and predictive criterion validity, respectively) and reported in the relevant Results sections.

Aim 2: Assess UBS Reliability. Two main approaches were utilized to examine the reliability of the UBS. First, internal consistency was evaluated by calculating Cronbach's alpha coefficients for each to-be-determined UBS subscale and total score at all four UBS quarterly administrations (i.e., Q1-Q4). Cronbach's alpha coefficients provide an estimation of whether the items on the total scale or individual subscales are measuring the same construct and doing so reliably. Cronbach's alpha coefficients were interpreted based on widely accepted guidelines (George & Mallery, 2003), indicating excellent reliability at $\alpha > .80$, good reliability at $\alpha > .70$, and acceptable reliability at $\alpha > .60$. Additionally, Nunnally and Bernstein's (1994) recommendation of $\alpha > .80$ for tools used in clinical practice was used as a guideline for an adequate demonstration of internal consistency reliability for UBS total and to-be-determined subscale scores. This was consistent with recommendations in the literature and psychometric investigations of other screening measures (Dvorsky et al., 2014; Lane, Oakes, et al., 2012; Martin & Savage-McGlynn, 2013).

Next, the stability of continuous scores across the four time points within the year were explored. Zero-order bivariate Pearson correlations were calculated for the UBS total and to-be-determined subscale scores between Q1, Q2, Q3, and Q4 UBS administrations to examine the temporal stability of UBS scores. The time intervals between adjacent administrations were as follows: Q1 to Q2 (2 months, 2 days), Q2 to Q3 (3 months, 1 day), and Q3 to Q4 (2 months, 9 days; S. Myers, personal communication, August 5, 2019). Time periods between non-adjacent administrations were: Q1 to Q3 (5 months, 3 days), Q2 to Q4 (5 months, 10 days), and Q1 to Q4 (7 months, 12 days). Cohen's (1988) guidelines for r will be used to interpret small ($r = .10$),

medium ($r = .30$), and large ($r = .50$) effect sizes. Given the unknown nature of UBS forthcoming subscale and total scores over time, the temporal stability in the current sample was also examined for students' BIMAS-2 scores across available administrations to serve as a relative basis of comparison for UBS temporal stability coefficients. In their good practice guide for psychometric evaluations, Martin and Savage-McGlynn (2013) note recommendations of $r \geq .70$ for acceptable test-retest reliability with time intervals of three months considered adequate (Kline, 2000); however, these are not specific to mental health screeners. Thus, this guideline was considered loosely for interpreting the adequacy of UBS test-retest coefficients.

Aim 3: Investigate UBS Convergent and Concurrent Criterion-Related Validity.

Convergent and concurrent criterion-related validity patterns of the UBS were examined by exploring the relationship between UBS subscales and total scale with BIMAS-2 subscale scores obtained at the same time point (i.e., Q2). Criterion validity in this case refers to the performance of the UBS relative to a reference standard, which in this study was the BIMAS-2. Convergent validity patterns were first explored between the UBS total scale and to-be-determined subscales and the BIMAS-2 subscales using zero-order Pearson bivariate correlations. Cohen's (1988) guidelines for r were used to interpret small ($r = .10$), medium ($r = .30$), and large ($r = .50$) effect sizes. Correlations were also examined for indications of redundancy between UBS total scale or to-be-determined subscales and BIMAS-2 subscales, with $r > |.70|$ suggesting potential redundancy between the two measures (Kline, 1979). Demonstrated effect sizes and statistical significance for correlation coefficients were then used to plan subsequent logistic regression analyses. If these strategies did not sufficiently discriminate relationships between subscales, Fisher's z -tests (Meng et al., 1992) were conducted to determine if the BIMAS-2 subscale under

examination correlated significantly more with the hypothesized UBS subscale than non-hypothesized subscale.

Following this set of analyses, concurrent criterion-related validity of the UBS was explored. Specifically, single-predictor logistic regression models were fitted to the data to test research hypotheses regarding the relationship between Q2 UBS total and to-be-determined subscale scores and a student's binary risk status based on Q2 BIMAS-2 subscale scores. This set of analyses were proposed to answer the question: *To what extent do scores on UBS subscales and total scale correctly classify students who were identified as any risk/concern on a given BIMAS-2 subscale at that same time point?* Thus, classification accuracy of the UBS was the focal point of this set of analyses, using BIMAS-2 subscale binary risk classification as the criterion or *reference standard* (Trevethan, 2017). Detailed information about the psychometric properties of the BIMAS-2, including description of classification accuracy metrics (e.g., sensitivity, specificity) in diagnostic samples, is provided in the description of measures.

While all model fit indices were examined for logistic regression analyses, I followed Hosmer and Lemeshow's (2000) suggestion to focus on the classification tables and related metrics produced by the logistic regression analyses as the primary metric for assessment of fit, particularly related to the validity of predicted probabilities in the model (as cited in Peng et al., 2002). Results of the classification power of the UBS focused on four main indices commonly used as determinants of the extent to which a screening measure can identify the likely presence or absence of a given condition: (a) sensitivity, (b) specificity, (c) positive predictive values (PPV), and (d) negative predictive values (NPV; Trevethan, 2017). These estimates were calculated based on the formulas displayed in Figure 1 using values provided in the classification tables produced by SPSS Version 27 (SPSS, 2020).

Figure 1

Equations for Calculating Sensitivity, Specificity, and Predictive Values

		Result from screening measure (i.e., UBS total scale/subscale score)		
		Positive (Any risk/concern status)	Negative (Low risk/no concern status)	
Status of Student according to reference standard (BIMAS-2 subscale risk status)	Has the condition (BIMAS-2 Any risk/concern status)	(a) True Positive	(c) False Negative	← Sensitivity [a / (a+c)] x 100
	Does not have the condition (BIMAS-2 Low risk/no concern status)	(b) False Positive	(d) True Negative	← Specificity [d / (d+b)] x 100
		↑ PPV [a / (a + b)] x 100	↑ NPV [d / (c + d)] x 100	

Note. Italicized text provides information specific to the current study regarding the condition the cell is describing. As the reader may notice, this predictor-criterion 2x2 table has been rotated from how it is traditionally displayed (i.e., criterion values in columns and predictor values in rows). This was transposed to correspond to SPSS Version 27 classification table results displayed in Appendices G through L. Table adapted from “Sensitivity, Specificity, and Predictive Values: Foundations, Plabilities, and Pitfalls in Research and Practice,” by R. Trevethan, 2017, *Frontiers in Public Health* 5(307), p. 2 (doi: 10.3389/fpubh.2017.00307).

A more detailed review of the definitions of sensitivity, specificity, and predictive values, as well as their relevance, can be found in the introduction section of this manuscript. For the current study, students were assigned to each of the four cells labeled (a) through (d) in Figure 1

based on whether they were identified as having or not having the target condition (i.e., at risk for having or developing social, emotional, or behavioral concerns) using the reference standard (i.e., *T*-scores in the *any risk* or *concern* range on a given BIMAS-2 subscale), and whether the screener (i.e., UBS subscales and/or total scale scores) yielded a positive result (i.e., the student appears to have the risk status) or a negative result (i.e., the student appears *not* to have the risk status; Trevethan, 2017).

For the current study, *sensitivity* focuses on the UBS's probability of identifying true positives while avoiding false negatives and was calculated from only those students who had positive results on the reference standard (i.e., BIMAS-2 *any risk/concern* status). *PPV* characterizes the UBS's probability of maximizing true positives while avoiding false positives and was calculated from only students who tested positive on the UBS as "having the condition" (i.e., predicting BIMAS-2 *any risk/concern* status). *Specificity* is concerned with the UBS's probability of identifying true negatives while avoiding false positives and was calculated based only on individuals who had negative results on the reference standard (i.e., BIMAS-2 *low risk/no concern* status). *NPV* characterizes the UBS's probability of identifying true negatives while avoiding false negatives and was calculated only from individuals who tested negative on the UBS for "having the condition" (i.e., predicting BIMAS-2 *low risk/no concern* status).

In contrast to other analytic indices (e.g., correlation coefficients), guidelines for interpreting the magnitude of the values obtained for sensitivity, specificity, PPV, and NPV are less clear. As described by Glover and Albers (2007) in their review of considerations for universal screening assessments, recommendations in the literature have noted 75% to 80% as minimum benchmarks for sensitivity, specificity, PPV, and NPV for a screening measure (Carran & Scott, 1992; Carter et al., 2004). In general, values closer to 100% for a given index

indicate the UBS is performing as well as the reference standard in terms of classification. However, the specific indices that are most important for evaluating a given screening measure depend on the priorities of that measure and the consequences of over-identification or under-identification. Regarding the UBS, more importance was placed on obtaining high sensitivity values with less interest in specificity values, since the UBS is used as a first gate of screening in a multi-gating system of identification. As such, false negatives (i.e., missing a student at risk for a given problem) are viewed as more detrimental than a false positive (i.e., incorrectly labeling a student as at risk of concerns; Glover & Albers, 2007). Regarding predictive values, high negative predictive values (NPVs) and low or moderate positive predictive values (PPVs) were preferred in the current study, as it is better to allow for more false positive risk status results (i.e., over-identification; high NPV) while minimizing false negative risk status results (i.e., under-identification; moderate PPVs) at the initial gate of screening.

The specific combinations of UBS subscales and/or total scale and BIMAS-2 subscales included in logistic regression analyses were determined by results of the Pearson correlations in concert with interpretations of the underlying constructs most closely represented by forthcoming UBS subscales. For example, if one of the UBS to-be-determined subscales seemed to conceptually relate to BIMAS-2 Conduct subscale and this was also supported by a significant correlation between those two subscales, then a logistic regression was conducted predicting BIMAS-2 Conduct subscale at risk status from said UBS subscale.

Aim 4: Explore UBS Short-Term Predictive Criterion-Related Validity. Short-term predictive criterion-related validity patterns of the UBS were investigated by examining the relationship between UBS subscale/total scale scores at Q1 and BIMAS-2 subscale binary risk status (the criterion) at Q4. This was specified as short-term predictive validity due to the

relatively short time interval (i.e., 7 months, 12 days) between Q1 UBS screener administration and Q4 BIMAS-2 criterion measure administration. Procedures for predictive validity analyses mirrored those conducted in Aim 3 for concurrent validity, including the same approach to correlations, logistic regressions, and classification accuracy.

First, zero-order Pearson correlations were conducted between Q1 UBS total and to-be-determined subscale scores and Q4 BIMAS-2 subscale scores using Cohen's (1988) guidelines to interpret the magnitude of correlations according to small ($r = .10$), medium ($r = .30$), and large ($r = .50$) effect sizes. The effect sizes and significance of correlation coefficients were then used to plan subsequent logistic regression analyses to pursue. If these did not sufficiently discriminate relationships between subscales, Fisher's z -tests (Meng et al., 1992) were conducted to determine if the BIMAS-2 subscale under examination correlated significantly more with the hypothesized UBS subscale than non-hypothesized subscale.

Results of these correlations were used in tandem with hypothesized relationships between UBS and BIMAS-2 subscales and/or total scale to guide logistic regression analyses. Separate simple logistic regression analyses were conducted with Q1 UBS total scale and to-be-determined subscales as the predictor and Q4 BIMAS-2 subscale binary risk or concern status as the outcome. This set of analyses were proposed to answer the question: *To what extent do scores on UBS subscales and total scale at the beginning of the school year correctly predict (i.e., classify) students who were later identified in the any risk/concern range on a given BIMAS-2 subscale at the end of the year?* The focus of these targeted logistic regression analyses was to examine the classification power of the UBS at the beginning of the academic year to predict end-of-year BIMAS-2 subscale risk or concern status (e.g., *low risk* v. *any risk*). In these

analyses, BIMAS-2 risk or concern status at Q4 was treated as the observed or actual classification.

Like Aim 3, in addition to examining model fit for the logistic regression analyses, focus was placed on examining aspects of classification accuracy for UBS total and subscale scores at Q1, including: sensitivity, specificity, PPV, and NPV values. This was because the ultimate purpose of Aim 4 was to evaluate the accuracy of UBS total and subscale scores at Q1 in predicting risk status classification at Q4 as determined by the criterion or *reference standard* (i.e., BIMAS-2 subscale scores). Goals for these adequacy metrics regarding predictive criterion validity were the same as described for concurrent criterion validity: (a) high sensitivity values (i.e., identifying true positives, minimizing false negatives); (b) high NPVs (i.e., allowing for more false positives); and (c) moderate PPVs (i.e., minimizing false negatives). As explained in the analytic plan for Aim 3, these characteristics are preferred for the initial gate of screening, when it is more detrimental to under-identify students than to over-identify students potentially at risk of having or developing social, emotional, or behavioral challenges (Glover & Albers, 2007).

Results

Data Diagnostics and Preparation

Two students were found to have the same unique student ID in the dataset despite having different demographic information. These data errors were resolved in consultation with one of the UBS developers who confirmed these two entries represented two separate students at the school and confirmed subsequent data were correctly entered. One of the students was given a new unique student ID and both students' data were retained in the dataset. No other data entry

errors were identified in the dataset, including impossible values. No other modifications were made to the raw data.

The UBS data were mostly complete, with 99.6% of the students identified as having no missing items across all four quarters. One student was found to have one missing item (Item 9) on the UBS at one time point (Q4). This student only had data available for Q3 and Q4, and thus could only be included in Aims 2 and 4, which both indicated sufficient sample size for analyses such that removal of this one participant would not impact power. Based on the combination of the very low scope of missingness and lack of impact on power, I elected to remove this participant from all analyses³. Thus, of the original 231 through sixth-grade students, 230 were retained for the final sample, all of whom had data for 100% of the items on the UBS and BIMAS-2 at the time point(s) at which any data were available.

As the reader may come to notice in the results below, across all four time points and all participants, there are a total of eight students at Q2 (i.e., 3.8% of Q2 administrations) and two students at Q4 (i.e., 0.9% of Q4 administrations) who have completed UBS data and no BIMAS-2 data. As previously noted, the HIDOE psychologists in charge of data collection reported the two measures were given at the same time and it remains unknown as to why these sample sizes are different. However, given the small nature of this discrepancy (i.e., 0.9% to 3.8% of the sample), I retained these cases in Aims 3 and 4 and utilized pairwise deletion strategies for relevant analyses (e.g., Pearson correlations, logistic regressions).

Examination of distributional properties of the data occurred in a stepwise fashion and varied slightly by study aim and measure. First, item-level descriptive statistics were examined

³ Correlational analyses conducted exploratorily with this participant included ($N = 231$) showed no qualitative differences in effect sizes of correlations with or without the participant with missing datum.

for UBS data across all time points to confirm which assessment occasion to use in the EFA for Aim 1. For Aims 2 – 4, distributional properties were examined at the total scale and subscale level and were conducted based on the results of Aim 1 regarding UBS factor structure.

Aim 1: Data Diagnostics and Preparation

Regarding the first aim, item-level descriptive statistics were examined across all four time points to confirm that the proposed time point for use in the EFA (i.e., Q2) was appropriate. The primary consideration was ensuring Q2 data were not outliers and thus, would produce EFA results representative of the true performance of the UBS. Item-level means, standard deviations, skew, and kurtosis were examined across all four time points (see Appendix E for Q1, Q3, and Q4; and Table 4 for Q2). Means were compared across items to assess for performance differences across time points. In general, the majority of item means increased at each time point, indicating improvements in functioning over time, and similar patterns between items were found at each time point (see Table 5). For example, at each time point, Item 9 was found to have the lowest mean score relative to the other items, with Item 6 having the highest. Q2 data were found to be the least skewed and least kurtotic compared to the other three time points. As shown in Table 4, skewness and kurtosis values for all UBS items at Q2 were within an acceptable range (i.e., values between -2 to 2; George & Mallery, 2003). At least one item was found to be unacceptably kurtotic at Q1 and Q3, and Q1 and Q4 both had one item with skewness value greater than |1|. Overall, no major concerns were identified for Q2 data, and there were no indications that factors determined from using Q2 data might be different from those extracted at a different time point. As such, Q2 data were utilized to conduct the EFA⁴.

⁴ As an exploratory analytic safeguard, additional EFAs were conducted using the same methodology as reported for Q2 with each of the remaining three time points. Results of these analyses indicated no substantive differences in factors extracted, variance explained by each factor, and factor loadings between the four time points.

Table 4

Q2 UBS Item-Level, Corrected Item-Total Correlations, and Cronbach's Alpha if Item Deleted Statistics (n = 213)

UBS Item	<i>M</i>	<i>SD</i>	Skew	Kurtosis	Item-total <i>r</i>	α
Factor 1: Social/Emotional Engagement	3.71	0.65	-0.77	0.60		.92
1. Cooperates with peers	3.68	0.77	-0.64	0.15	.79	.90
2. Uses socially appropriate responses	3.62	0.78	-0.74	0.36	.85	.88
6. Has a positive attitude	3.78	0.69	-0.62	0.66	.76	.90
7. Regulates emotions	3.77	0.78	-0.72	0.38	.76	.90
5. Follows rules, routines, & directions	3.72	0.72	-0.42	0.14	.76	.90
Factor 2: Academic Readiness	3.52	0.75	-0.39	-0.10		.92
3. Is prepared to learn	3.69	0.81	-0.18	-0.43	.81	.90
4. Engages in academic tasks	3.57	0.84	-0.40	-0.22	.82	.90
8. Is an effective problem solver	3.42	0.86	-0.47	0.04	.74	.91
9. Pays attention	3.39	0.86	-0.44	0.16	.74	.89
5. Follows rules, routines, & directions	3.72	0.72	-0.42	0.14	.86	.91
Total Scale (All 9 items)						.94
1. Cooperates with peers					.75	.94
2. Uses socially appropriate responses					.83	.93
3. Is prepared to learn					.79	.93
4. Engages in academic tasks					.77	.94
5. Follows rules, routines, & directions					.80	.93
6. Has a positive attitude					.75	.94
7. Regulates emotions					.73	.94
8. Is an effective problem solver					.74	.94
9. Pays attention					.84	.93

Note. The Cronbach's alpha coefficients reported at item-level are *alpha if item removed* values, whereas alpha coefficients reported for each factor are the factor-level Cronbach's alpha values. Item 5 is included in both subscales to report item-total statistics relevant to retainment on either factor. This item is bolded to emphasize these results. Item-total statistics suggest a greater reduction in alpha if Item 5 is deleted from Factor 1 than if deleted from Factor 2. However, item-total correlation is higher for Item 5 on Factor 2 than on Factor 1.

Table 5*UBS Item-Level Means and Standard Deviations for Q1 – Q4*

UBS Item	Q1 <i>M(SD)</i>	Q2 <i>M (SD)</i>	Q3 <i>M (SD)</i>	Q4 <i>M (SD)</i>
1	3.48 (0.81)	3.68 (0.77)	3.74 (0.76)	3.87 (0.82)
2	3.42 (0.84)	3.62 (0.78)	3.62 (0.78)	3.83 (0.80)
3	3.40 (0.83)	3.69 (0.81)	3.62 (0.82)	3.79 (0.85)
4	3.26 (0.89)	3.57 (0.84)	3.69 (0.79)	3.83 (0.85)
5	3.55 (0.73)	3.72 (0.72)	3.72 (0.77)	3.86 (0.83)
6	3.59 (0.77)	3.78 (0.69)	3.84 (0.64)	4.04 (0.74)
7	3.51 (0.83)	3.77 (0.78)	3.81 (0.76)	3.92 (0.90)
8	3.23 (0.82)	3.42 (0.86)	3.50 (0.79)	3.68 (0.82)
9	3.08 (0.96)	3.39 (0.86)	3.48 (0.82)	3.67 (0.89)

Note. UBS scores range from 1.0 to 5.0 for each item.

Aims 2 Through 4: Data Diagnostics and Preparation

Examination of distributional properties were relevant for correlation analyses in Aims 2, 3, and 4, and Cronbach's alpha coefficient estimates in Aim 2. Across analyses UBS data across all four time points were used, as well as both available time points of BIMAS-2 data. Thus, for the second through fourth aims, UBS subscales⁵, UBS total scale, and BIMAS-2 subscales minimum and maximum values, means, standard deviations, Shapiro-Wilk's W statistics, skewness, kurtosis, and number of statistical outliers were examined (see Table 6 for UBS and Table 7 for BIMAS-2 results). Results are presented separately for each total scale and/or subscale and time point (i.e., quarter) of measure administration.

⁵ As the results in Aim 1 will indicate, results of the EFA revealed a two-factor model for the UBS. Thus, study results from here on will focus on the two-factor model and include examinations of UBS Factor 1: Social/Emotional Engagement, UBS factor 2: Academic Readiness, and UBS total scale scores.

Regarding the UBS, data for subscales and total scale scores across all time points showed skewness in the “excellent” (-1.0 to 1.0) and kurtosis in the “excellent” to “acceptable” (-2.0 to 2.0) range (George & Mallery, 2003). Given the absence of excessive skewness or kurtosis in the sample, the threshold for a large enough sample to assume normality was likely met by the current sample. The smallest subset of the sample used in analyses across all aims was $n = 198$, which is greater than the 160 participants reported by Field (2018) as a general guideline for the central limit theorem. Shapiro-Wilk’s W statistic was significant for both UBS subscales and total scale at each time point, suggesting significant deviation from normal distribution. However, this should be interpreted with caution due to the potential influence of large samples on this test statistic. Regarding outliers, examination of box plots suggested several potential outliers across total scale/subscales and time points. Outliers were examined and none were found to reflect errors in the data. Since outliers potentially represented cases reflecting more concerns/risk than found in the majority population (i.e., students the measure is designed to identify) and thus, could have represented critical information for the current study, they were ultimately retained. The scope and distribution of outliers were further explored through z-scores to ensure proposed parametric tests in Aims 2-4 were appropriated to conduct. A detailed summary of this analysis can be found in Appendix F and Table F1. Z-score assessment of normality and outliers of UBS subscales and total scale presented mixed results, indicating some minor concerns with outliers, particularly related to the less than 95% of cases falling in the normal range for scores at Q2 and Q3. However, the divergence from the expected 95% of cases in the normal range appeared to be low, with 4.4% fewer cases in the normal range for Q2 Factor 1, 3.5% fewer cases for Q2 Factor 2, and 3.8% fewer cases for Q4 Factor 2 and total scale than the 95% benchmark of a normal distribution. Regarding range of scores, Q1 and

Q3 scale scores utilized the full range of values, compared with Q2 and Q4 which showed some restricted range in minimum values (see Table 6). This means for these subscales and total scales, no students were rated the lowest possible score (i.e., 1; lower scores on the UBS indicate poorer functioning).

Table 6

Descriptive and Normality Statistics for Q1 – Q4 UBS Subscale and Total Scale Scores

UBS Scale	<i>M</i>	<i>SD</i>	Min	Max	Shapiro -Wilk's <i>W</i>	Skew	Kurtosis	% Valid Cases in Normal Range ^a
Q1 (<i>n</i> = 214)								
UBS Factor 1	3.51	0.72	1.00	5.00	.92*	-0.96	1.70	94.9%
UBS Factor 2	3.24	0.78	1.00	5.00	.96*	-0.54	0.15	95.3%
UBS Total	3.39	0.71	1.00	5.00	.95*	-0.80	1.18	94.9%
Q2 (<i>n</i> = 213)								
UBS Factor 1	3.71	0.65	1.80	5.00	.92*	-0.77	0.60	90.6%
UBS Factor 2	3.52	0.75	1.50	5.00	.97*	-0.39	-0.10	91.5%
UBS Total	3.62	0.65	1.67	5.00	.97*	-0.55	0.21	94.8%
Q3 (<i>n</i> = 218)								
UBS Factor 1	3.75	0.65	1.00	5.00	.90*	-0.81	1.70	95.4%
UBS Factor 2	3.57	0.72	1.00	5.00	.94*	-0.48	0.45	91.2%
UBS Total	3.67	0.65	1.00	5.00	.94*	-0.59	1.34	91.2%
Q4 (<i>n</i> = 214)								
UBS Factor 1	3.90	0.73	1.40	5.00	.94*	-0.60	0.52	95.3%
UBS Factor 2	3.74	0.76	1.00	5.00	.96*	-0.66	0.67	95.8%
UBS Total	3.83	0.71	1.44	5.00	.97*	-0.57	0.62	96.2%

^a Percentage of valid cases in normal range used as assessment of outliers. Approximately 95% of cases in normal range represents normal curve, thus, it is desirable that percentages are at or around 95% (Field, 2018).

* $p < .001$.

Similar findings are reported for BIMAS-2 subscales regarding skewness, kurtosis, and statistical tests of normality (see Table 7). Specifically, all BIMAS-2 subscale scores were found to have skewness and kurtosis in the “excellent” to “acceptable” range across both time points (George & Mallery, 2003). The Shapiro-Wilk’s W statistic of normality was found to be significant for all subscales at both time points. Regarding outliers, examination of box plots suggested several potential outliers across subscales and time points.

Table 7

Descriptive and Normality Statistics for Q2 and Q4 BIMAS-2 Subscale T-scores

BIMAS-2 Subscale	<i>M</i>	<i>SD</i>	Min	Max	Shapiro- Wilk’s <i>W</i>	Skew	Kurtosis	% Valid Cases in Normal Range
Q2 (<i>n</i> = 205)								
Conduct	52.13	7.25	43	76	.92*	0.91	0.67	95.6%
Negative Affect	52.45	10.17	40	80	.94*	0.56	-0.45	95.6%
Cognitive/Attention	52.81	11.90	31	82	.97*	0.41	-0.36	97.1%
Social	49.22	10.59	23	73	.98*	0.17	-0.05	91.7%
Academic Functioning	49.11	10.09	20	68	.98*	-0.44	0.18	97.1%
Q4 (<i>n</i> = 211)								
Conduct	51.56	7.26	43	81	.90*	1.16	1.71	96.7%
Negative Affect	49.67	8.59	40	79	.91*	0.73	-0.02	97.2%
Cognitive/Attention	50.27	11.96	31	80	.97*	0.40	-0.57	95.3%
Social	52.74	9.82	28	73	.95*	0.34	-0.06	97.6%
Academic Functioning	51.49	10.70	21	68	.97*	-0.37	-0.17	97.2%

Note. Bolded values for min and max indicate values below 30 or above 70, highlighting subscales with values considered unusually low and unusually high. Percentage of valid cases in the normal range were used as assessment of outliers. Approximately 95% of cases in the normal range represents normal curve, thus, percentages at or around 95% are desirable (Field, 2018).

* $p < .001$

A closer examination of the scope and distribution of BIMAS-2 outliers using z -scores can be found in Appendix F and Table F2. Overall, this examination revealed these outliers only minimally diverged from expected ranges of a normal distribution. Regarding distribution of BIMAS-2 scores, T -scores range from 0 to 100 with a mean of 50 and standard deviation of 10. If the scores are normally distributed, it is expected that approximately two-thirds of scores will fall between 40 and 60 and approximately 95% of scores will fall between 30 and 70. As shown in Table 7, means and standard deviations for subscales at both time points generally fall within a few points of $M = 50$ and $SD = 10$.

Overall, data for both UBS and BIMAS-2 were deemed not heavily skewed or kurtotic across all subscales and/or total scale at all time points. Given the large sample size, less weight was given to the significant Shapiro-Wilk's W statistics found for subscales and/or total scale on both measures across all time points. Due to minimal issues with non-normality and the modest number of outliers, all proposed parametric analyses across Aims 2-4 were conducted as planned. Specifically, Cronbach's alpha coefficients conducted in Aim 2 and Pearson bivariate correlations used in Aims 2 – 4 were utilized as proposed, as they were likely robust to the minimal level of non-normality reported in the current data (Bishara & Hittner, 2012; Field, 2018). Logistic regressions in Aims 3 and 4 were also conducted as proposed, given they do not rely on distributional assumptions (Field, 2018).

Examining the Influence of Gender on BIMAS-2 Scores. As proposed, t -tests were conducted to explore the influence of gender on student BIMAS-2 T -scores in preparation for planned logistic regression analyses in Aims 3 and 4. Gender was only found to significantly relate to students' BIMAS-2 Cognitive/Attention T -scores at both Q2 and Q4. Specifically, at Q2, the 83 female students ($M = 49.96$, $SD = 11.72$) compared to the 118 male students ($M =$

54.67, $SD = 11.75$) evidenced significantly lower scores on the BIMAS-2 Cognitive/Attention subscale, $t(199) = 2.80, p = .006$. Similarly, at Q4, the 89 female students ($M = 47.54, SD = 11.52$) compared to the 122 male students ($M = 52.21, SD = 11.98$) evidenced significantly lower scores on the BIMAS-2 Cognitive/Attention subscale, $t(209) = 2.85, p = .005$.

Aim 1: UBS Factor Structure

As planned, IBM SPSS Statistics Version 23 (SPSS, 2015) was used to conduct the exploratory factor analysis (EFA) on the nine-item UBS to understand its underlying factor structure. Since findings from preliminary examination of outliers suggested a mild degree of nonnormality for the UBS total scale at Q2, the principal axis factoring method was utilized to be cautious as it does not require distributional assumptions (Fabrigar et al., 1999)⁶. Oblique (direct oblimin) rotation was utilized to account for the likelihood that factors would be correlated.

Several analyses were explored to examine assumptions, factorability, and sampling adequacy. First, the factorability of the nine UBS items was examined through a series of analyses. Bivariate Pearson correlations between all items were utilized to assess for potential issues of multicollinearity (i.e., correlations that are too high; $r > 0.8$) or poor fit with the larger item pool (i.e., correlations that are too low; $r < 0.3$; Field, 2018). As can be seen in Table 8, all correlations fell above $r = 0.4$ and at or below $r = 0.8$, suggesting no major issues with multicollinearity or a lack of patterned relationships amongst items (Yong & Pearce, 2013).

⁶ An exploratory EFA using maximum likelihood method was conducted post hoc and revealed no substantial differences in any metric, including factor structure and strength of factor loadings across items. Furthermore, the principal axis factor analysis was conducted exploratorily using promax rotation instead of direct oblimin and there were also no substantial differences across any metrics between the two versions. All post hoc alternative methods explored for different approaches and rotations produced the same factor structure and similar loadings as the current analysis.

Table 8*Pearson Bivariate Correlations Between UBS Items 1-9 (n = 213)*

	1	2	3	4	5	6	7	8	9
1. Cooperates with peers	--								
2. Uses socially appropriate responses	.80	--							
3. Is prepared to learn	.57	.64	--						
4. Engages in academic tasks	.52	.61	.79	--					
5. Follows rules, routines, & directions	.72	.73	.68	.65	--				
6. Has a positive attitude	.61	.69	.57	.58	.62	--			
7. Regulates emotions	.62	.70	.60	.49	.61	.74	--		
8. Is an effective problem solver	.54	.61	.64	.67	.56	.57	.56	--	
9. Pays attention	.64	.69	.73	.77	.75	.60	.58	.74	--

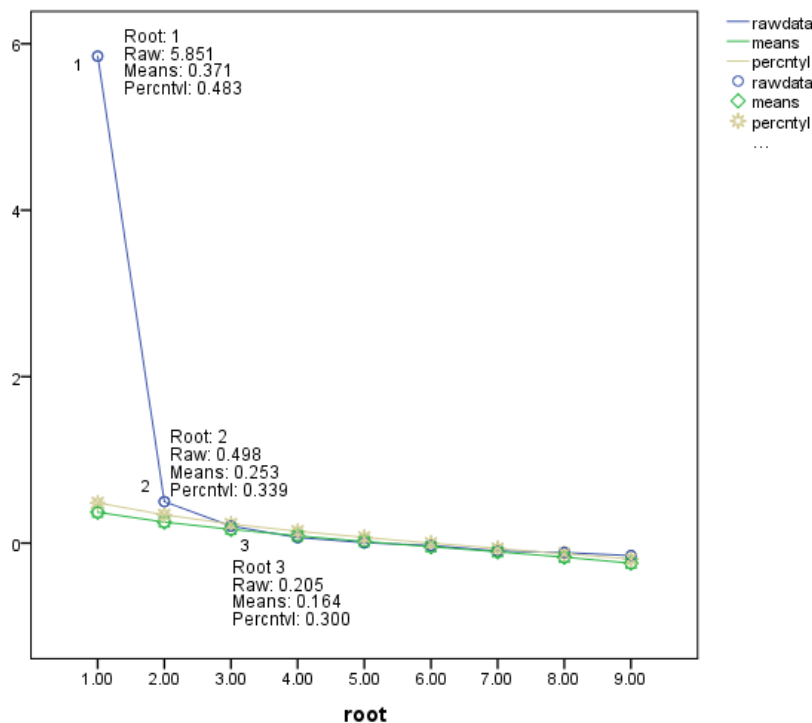
Note. All correlations were significant at $p < .001$.

Next, sampling adequacy was examined. First, the minimum amount of data for factor analysis was satisfied by Q2 UBS data with a final sample of 213 students, providing a ratio of over 23 participants per variable, which is well above common 5:1 or 10:1 ratio recommendation (Yong & Pearce, 2013). Examination of factor loadings confirmed the adequacy of the sample size with eight of nine variables achieved loadings $> .60$, indicating a minimum sample of 150 as sufficient (Guadagnoli & Velicer, 1988). The Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy for the Q2 sample appeared adequate ($KMO = 0.91$) falling into the ‘marvelous’ range according to Kaiser and Rice’s (1974) guidelines. KMO values for individual items ranged from 0.88 to 0.93, which is well above the acceptable limit of 0.5 (Kaiser & Rice, 1974). Furthermore, Bartlett’s test of sphericity was significant ($\chi^2 [36] = 1595.39, p < .001$) indicating correlations between items were significantly different than zero. Given these overall indicators, factor analysis was deemed appropriate to use with all nine items on the UBS and likely to yield distinct and reliable factors (Field, 2018).

Prior to conducting the factor analysis, a parallel analysis was conducted using 1,000 permutations of the raw data set (O'Connor, 2000). As seen in Figure 2, parallel analysis at the 95th percentile eigenvalue suggested up to two factors be retained. An initial principal-axis factor analysis with oblique (direct oblimin) rotation was conducted examining the two-factor solution. Results indicated the first and second factors explained 68.39% and 8.91% of the variance respectively (see Figure 3 for eigenvalues) with a combined 77.29% of variance explained.

Figure 2

Parallel Analysis Scree Plot

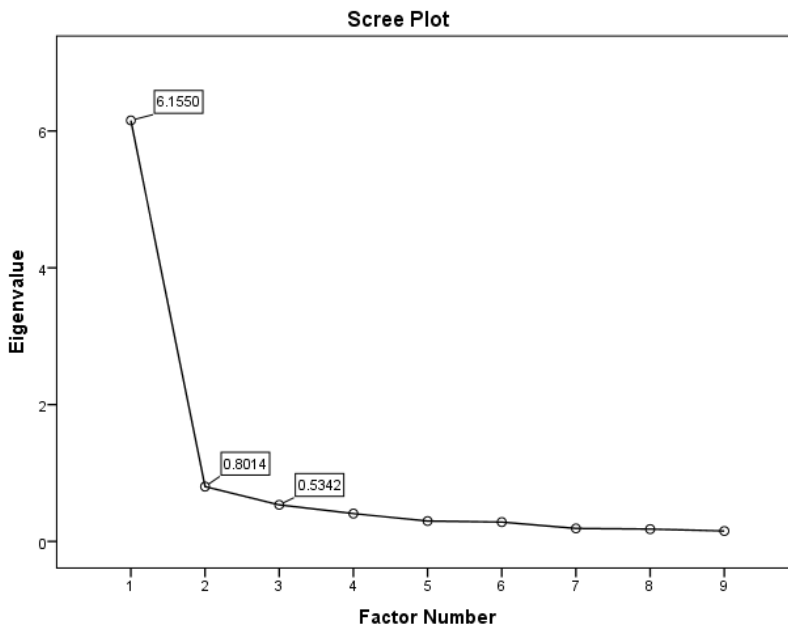


Note. Eigenvalues are statistically significant if the rawdata eigenvalue exceeds that of the percentyl. As seen above, values reported for the first and second factors meet this criterion, whereas the third factor does not.

Multiple metrics were examined to determine the appropriateness of a two-factor solution, including scree plots and eigenvalues. Consistent with the parallel analysis, the scree plot produced by the EFA suggested a potential point of inflection justifying retaining a maximum of two factors (Figure 3). Regarding eigenvalues, Kaiser's criterion of 1 (Kaiser, 1960, 1970), was not met for the second factor (eigenvalue = 0.80), although the more liberal Joliffe criterion (i.e., eigenvalue of 0.7 needed for retention; Joliffe, 1972, 1986) was achieved. According to Field (2018), Kaiser's criterion can be assumed accurate when "there are less than 30 variables and communalities after extraction are greater than 0.7" (p. 811). The current data meets the variable guideline, but only four items evidenced communalities after extraction greater than 0.7; the remaining five items had communalities ranging from .60 to .68. Taken together, the two-factor solution, which explained 77.29% of the variance, was preferred because: (a) the parallel analysis and scree plot support a two-factor model, (b) the eigenvalue for the second factor meets the more liberal Joliffe's criterion, and (c) there is some evidence suggesting Kaiser's criterion (i.e., the only result not in support of a two-factor model) may not be assumed accurate for the data.

Figure 3

Scree Plot Produced by Factor Analysis



Note. Scree plot produced for initial eigenvalues (before rotation) by common factor analysis using principal axis factoring method with original raw data. Initial eigenvalues reported for first three factors.

In interpreting the rotated factor pattern, an item was considered as loading onto a given factor if the factor loading was $\geq |.35|$ (Floyd & Widaman, 1995). As seen in Table 9, five items met this criterion for the first factor (i.e., 1, 2, 5, 6, 7) and four items for the second factor (3, 4, 8, 9). Across all items but one, there was clear differentiation in loadings for the two factors (e.g., Item 1 has loading $> |.80|$ on Factor 1 with loading $< |.1|$ on Factor 2) indicating no issues with cross loading. This item (Item 5) was found to load onto Factor 1 at .55 and Factor 2 at -.33. Although the -.33 loading on Factor 2 is below the $\geq |.35|$ criterion, it does not demonstrate a simple structure (i.e., relatively high loadings for one factor and low or near zero for the other),

which is important for interpretability (Lane, Oakes, et al., 2012). Thus, Item 5 was subjected to further evaluation. Prior to evaluating the cross-loaded item, the four items clearly loading onto each factor were further examined based on theoretical considerations (i.e., the extent to which items share conceptual meaning, the extent to which retained items between factors suggest distinct and different constructs, and whether there were at least three items on each factor with significant loadings). Placement of Item 5 was then based on its fit with the finalized items on the factors. The themes described below for both Factor 1 and Factor 2 were identified through close inspection of the items independently as well as through multiple collaborative discussions with my primary doctoral advisor and the two UBS measure developers (i.e., school and clinical psychologists within HDOE).

Table 9*Exploratory Factor Analysis: Factor Loadings for the UBS Two-Factor Model (n = 213)*

Item #	Item description	Factor 1	Factor 2	h^2
Factor 1: Social/Emotional Engagement				
2	Uses socially appropriate responses: <i>Responds appropriately to the emotions/behaviors of others, is an effective communicator</i>	.886	-.021	.82
1	Cooperates with peers: <i>Plays and works well with others, is kind, is a community contributor</i>	.839	.017	.68
7	Regulates emotions: <i>Does not demonstrate intense feelings of sadness, worry, anger, etc.</i>	.836	.037	.65
6	Has a positive attitude: <i>Demonstrates a growth mindset, is interested in improving the school community</i>	.738	-.073	.63
5	Follows rules, routines, and directions: <i>Responds safely to expectations and/or changes in the environment</i>	.546	-.329	.69
Factor 2: Academic Readiness				
4	Engages in academic tasks: <i>Starts, works, and finishes academic tasks within reasonable time frames, and is a quality producer</i>	-.147	-1.020	.83
3	Is prepared to learn: <i>Arrives on time, is responsible for school materials, is a self-directed learner</i>	.116	-.754	.72
9	Pays attention: <i>Is focused and is not overly distracted</i>	.184	-.740	.79
8	Is an effective problem solver: <i>Puts thoughts into decisions (i.e., not impulsive), is a complex thinker, and an ethical user of technology</i>	.199	-.607	.60

Note. Bolded values indicate loadings $> |.35|$ for the factor on which the item was retained. The extraction method was principal axis factoring with an oblique (direct oblimin method) rotation. Factor loadings and communality estimates (h^2) are included for each item. Communality estimates range from 0-1, with 0 indicating no shared variance between the item and the factor. No items were reverse scored.

As seen in Table 9, the items loading exclusively on the first factor were Items 1, 2, 6, and 7. These items can be characterized in several ways. First, these items appear to share conceptual meaning. Specifically, the items focus on the student's interaction with their social environment (e.g., response to others' emotions, contribution to the school community, plays and works well with others) or on prerequisites for interacting safely with their environment (i.e., emotional regulation). Second, these items all seem to have more of an external, observable component about the student's behavior. Third, these items appear to relate to behaviors that are observable across all school environments (i.e., inside and outside of the classroom). Lastly, all four items relate to social skills most directly, with some aspects of emotional regulation included as well.

Items loading clearly on the second factor (i.e., Items 3, 4, 8, and 9) also appear to share conceptual meaning that is distinct and different from that described for Factor 1. First, as shown in Table 9, items loading on Factor 2 appear to focus on being a responsible learner (e.g., engaging in tasks, preparedness for learning, paying attention, and problem solving). In contrast to Factor 1's focus on the student's interaction with social environment, items on Factor 2 seem more focused on the individual's behavior related to classroom performance (e.g., academic tasks, executive functioning) and how a student positions themselves to accept academic instruction. From a theoretical standpoint, these items also appear to relate to several of the symptoms of attention-deficit/hyperactivity disorder (AD/HD), including both inattention (e.g., starts and finishes tasks in timely manner, quality work production/attention to detail, distractibility, difficulty focusing, responsible for school materials [disorganization and forgetfulness]) and hyperactivity (e.g., impulsivity). Lastly, in contrast to Factor 1, items on

Factor 2 seem to relate to behaviors that are primarily observable in the classroom, with less emphasis on behavior outside the classroom setting.

The strong conceptual meaning shared between items on each factor, the clear differentiation between items loading onto each factor, and the presence of at least three items per factor, provided support for retaining all four strongly loading items on each factor. No items were considered for deletion. Following this, Item 5 was considered for placement on one of the two factors. The following parameters were considered in finalizing the placement of the cross-loaded item: (a) factor loading values of the item and (b) alignment with the other items and incremental benefit of including Item 5 on the factor. Regarding factor loading values, Lane, Oakes, and colleagues (2012) suggest a factor can be considered to load on a given factor if the loading is $\geq .40$ on that factor and $< .40$ on the remaining factor. Based on this criterion, Item 5 loaded onto Factor 1 and not Factor 2. Next, alignment with the other items on each factor was considered, as well as whether Item 5 provided any unique or incremental benefit when included on a given factor.

Item 5 (i.e., *Follows rules, routines, and directions: Responds safely to expectations and/or changes in the environment*) could be argued as adding a relevant contribution to both factors depending on which part of the item description is highlighted. Following rules, routines, and directions are important for both emotional regulation and social interaction as well as academic readiness and engagement, and classroom rules and routines are often relevant to both areas of student functioning. “Responding safely” to changes in the environment seems to have implications for emotional and social regulation as that is a prerequisite for safe responding in some situations (e.g., transitions require students to manage any strong positive or negative emotions associated with ending or starting a task to able to safely transition to the next activity).

Furthermore, this interpretation of the item appears aligned with other aspects of the first factor, such as the focus on the nature of students' interactions with their environment. Regarding alignment with Factor 2, classroom rules, routines, and directions are often given related to engagement in academic activities (e.g., stay on task, expectations to have homework or supplies, directing focus on the lesson). However, between the focus of Factor 1 on the student's social and emotional interaction with their larger school environment compared to the apparent focus of Factor 2 on the student's engagement on academic tasks within the classroom, Item 5 appears to be slightly more aligned with the focus on the student's interaction with the environment.

Despite factor loadings and other considerations suggesting that Item 5 load onto Factor 1, due to the lack of high clarity regarding this issue, two additional considerations were made for determining factor placement. First, reliability coefficients were examined for each factor with and without Item 5 to determine the impact of inclusion. Second, additional principal axis factor analyses using oblique (direct oblimin) rotation were conducted for each of the other three time points to examine consistency of factor loadings across available assessment occasions. Regarding Cronbach's alpha coefficients, there were no substantial differences in qualitative interpretation of coefficients with or without Item 5 for both factors. Specifically, Cronbach's alpha coefficients for each factor at all four time points improved slightly with the addition of Item 5 but remained in the excellent range at or above $\alpha = .90$ (George & Mallery, 2003), with the largest observed difference occurring at Q2 for Factor 1 (i.e., α with Item 5 = .915, α without Item 5 = .900). Regarding consistency in factor loadings for Item 5 across time points, loadings were either fairly even on both factors (i.e., Q1) or Item 5 loaded more strongly onto Factor 1 (i.e., Q2, Q3, and Q4). Taken together, results from these four additional queries of Item 5

suggested support for retaining the item on the first factor. While some examinations suggested similar support for Item 5 loading on either Factor 1 or Factor 2 (e.g., EFA results at other time points, conceptual alignment with other items on the factor, impact on reliability coefficients) other examinations showed support for Factor 1 (factor loadings of Item 5 at Q2, Q3, and Q4), and no examination showed preferential support for Factor 2 over Factor 1.

Following finalization of the two factors and their subsequent items, the two factors were named. This involved a multi-step process involving at least two individual meetings with my primary doctoral advisor and virtual discussions with the two UBS measure developers. A few recommendations for naming factors were considered, including: (a) identifying terms or phrases that best describe the items on a given factor (Yong & Pearce, 2013) and (b) focusing on names that capture the underlying latent and unobserved constructs behind those items (Henson & Roberts, 2006). Given their history with creating the items for the UBS, the original developers were instrumental in providing input about the underlying constructs items were intended to measure. Overall, the measure developers noted their emphasis on social and emotional learning goals (e.g., self-awareness, self-management, social awareness, etc.; CASEL, 2015; Denham, 2005) in the development of many of the items loading onto Factor 1 and their focus on HIDEOE general learning objectives (GLOs) for several of the items loading onto Factor 2 (e.g., being a self-directed learner, complex thinker, quality producer, ethical user of technology; S. Myers & K. McDonald, personal communication, March 11, 2021). Based on these discussions and considerations of the previously identified themes recognized across items (see discussion of initial four items retained on each factor), Factor 1 was named *Social/Emotional Engagement* to highlight the correspondence with broader SEL competencies observed across school environments (Zins et al., 2007; see also Denham, 2005). Factor 2 was named *Academic*

Readiness to highlight the focus on academic preparedness and learning behaviors generally observed within the classroom setting. The two-factor UBS described above, containing a total of 9 items, was determined to be the most parsimonious solution, and was utilized in subsequent analyses (i.e., reliability, convergent validity, and predictive validity). The total scale, including all nine UBS items, was also maintained along with the two newly established subscales, and named *UBS Total*. Scores on the UBS Total and both subscales range from 1.0 to 5.0 and are calculated as the average of scores across all items included on that particular subscale or total scale. Across UBS Total and subscales, higher scale scores indicate a greater degree of adaptive functioning. For Q2, the time point utilized in the EFA, the mean of the first factor of Social/Emotional Engagement was 3.71 ($SD = 0.65$), which was slightly higher than the mean of the second factor of Academic Readiness ($M = 3.52$, $SD = 0.75$). The final two subscales were significantly correlated at $r = .783$ ($p < .001$), which is characterized as a large effect size based on Cohens' (1988) guidelines.

Aim 2: UBS Reliability

Regarding Aim 2, examining the reliability of the UBS Total and subscale scores, a few tentative hypotheses were proposed. Based on visual inspection of the factors and items, as well as preliminary item-level analyses conducted for Aim 1, I anticipated finding Cronbach's alpha coefficients for UBS Total and subscales in the good to excellent range across all administrations (i.e., $\alpha > .70$; George & Mallery, 2003). In terms of test-retest reliability, I hypothesized stronger correlations for adjacent time points, with significant, but smaller correlation coefficients for comparisons with longer time intervals.

Cronbach's Alpha Coefficients

Cronbach's alpha coefficients were calculated for UBS Total and subscale scores at all four time points as a measure of internal consistency of items. Across all administrations, Cronbach's alpha coefficients for items on the UBS subscales and total scale fell in the excellent range ($\alpha > .80$; George & Mallery, 2003), consistent with hypotheses. Specifically, across all time points, the UBS Social/Emotional Engagement subscale ranged from $\alpha = .92$ to $.94$, the UBS Academic Readiness subscale ranged from $\alpha = .91$ to $.92$, and the nine-item total scale ranged from $\alpha = .94$ to $.95$ (see Table 10).

Test-Retest Correlation Coefficients

Test-retest reliability was also examined for the UBS Total and subscale scores across the four time points. Zero-order Bivariate Pearson Correlations were conducted between Q1, Q2, Q3, and Q4 UBS administrations. Time intervals corresponding to each comparison (i.e., Q1 – Q4 UBS scores) are included in Table 11 along with the correlations for the UBS Social/Emotional Engagement subscale scores across the four time points. Table 12 presents the correlation matrix for UBS Academic Readiness subscale and Table 13 presents the correlations for the UBS Total scores across time points. As seen in the tables, correlations between all four time points for each UBS subscale and Total scale evidenced a large effect size ($r > .50$; Cohen, 1988). As hypothesized, for each UBS subscale and Total scale, the correlations were strongest between adjacent time points (e.g., Q1 and Q2) and became weaker the larger the amount of time between administrations (i.e., Q1 and Q3 correlations weaker than Q1 and Q2, but stronger than Q1 and Q4 correlations). Test-retest correlation coefficients between Q2 and Q4 UBS scores (i.e., r ranging from $.66$ to $.70$) were similar to that found with the current sample for BIMAS-2 subscales over the same 5 months, 10 day time period (i.e., Q2 to Q4; $n = 200$): Conduct ($r = .63$;

$p < .001$), Negative Affect ($r = .66$; $p < .001$), Cognitive/Attention ($r = .79$; $p < .001$), Social ($r = .50$; $p < .001$); Academic Functioning ($r = .70$; $p < .001$).

Table 10

Cronbach's Alpha Coefficients for UBS Total and Subscale Scores for Q1 – Q4

Scale	No. of Items	Q1 α ($n = 214$)	Q2 α ($n = 213$)	Q3 α ($n = 217$)	Q4 α ($n = 213$)
UBS Factor 1	5	.94	.92	.92	.94
UBS Factor 2	4	.91	.91	.92	.91
UBS Total	9	.95	.94	.95	.95

Table 11

UBS Social/Emotional Engagement Subscale: Test-Retest Bivariate Correlations Between Scores at Each Time Point

Time Point	1	2	3	4
Q1	--	2 months, 2 days	5 months, 3 days	7 months, 12 days
Q2	.746 ($n = 207$)	--	3 months, 1 day	5 months, 10 days
Q3	.655 ($n = 205$)	.799 ($n = 207$)	--	2 months, 9 days
Q4	.583 ($n = 199$)	.706 ($n = 201$)	.701 ($n = 211$)	--

Note. All correlations significant at $p < .001$. Correlation coefficients and corresponding sample size are shown below the diagonal; test-retest time periods for each comparison are shown above the diagonal.

Table 12

UBS Academic Readiness Subscale: Test-Retest Bivariate Correlations Between Scores at Each Time Point

Time Point	1	2	3	4
Q1	--	<i>n</i> = 207	<i>n</i> = 205	<i>n</i> = 199
Q2	.756	--	<i>n</i> = 207	<i>n</i> = 201
Q3	.701	.824	--	<i>n</i> = 211
Q4	.656	.749	.767	--

Note. All correlations significant at $p < .001$. Correlation coefficients are shown below the diagonal; sample sizes are shown above the diagonal.

Table 13

UBS Total: Test-Retest Bivariate Correlations Between Scores at Each Time Point

Time Point	1	2	3	4
Q1	--	<i>n</i> = 207	<i>n</i> = 205	<i>n</i> = 199
Q2	.776	--	<i>n</i> = 207	<i>n</i> = 201
Q3	.696	.830	--	<i>n</i> = 211
Q4	.628	.743	.748	--

Note. All correlations significant at $p < .001$. Correlation coefficients are shown below the diagonal; sample sizes are shown above the diagonal.

Aim 3: UBS Convergent and Concurrent Criterion-Related Validity

Prior to exploring the criterion-related validity of the UBS scales, general hypotheses were made based on results of Aims 1 and 2. Broadly speaking, in terms of the anticipated direction of all associations, I suspected that all UBS scales would correlate in a negative

direction with BIMAS-2 Behavioral Concern Scales, as higher scores indicate more concerns on these BIMAS-2 scales while lower scores indicate more concerns on the UBS. For BIMAS-2 Adaptive Scales, I hypothesized positive correlations between the two measures, since these combinations of subscales are all strengths-based, with higher scores indicating better functioning. Regarding specific relationships between UBS subscales and BIMAS-2 subscales⁷, I speculated that zero-order Pearson bivariate correlation coefficients would be the strongest between pairwise comparisons for UBS Academic Readiness subscale and both the BIMAS-2 Academic Functioning and Cognitive/Attention subscales, given the seeming overlap in academic- and learning-related constructs. I suspected Academic Readiness would relate more strongly to these BIMAS-2 subscales than with BIMAS-2 Conduct, Social, and Negative Affect subscales given there was comparatively less overlap in item content between these subscales. Related to criterion validity patterns, I suspected that UBS Academic Readiness would show relatively better classification accuracy performance in classifying risk status for BIMAS-2 Academic Functioning and Cognitive/Attention subscales (compared to other BIMAS-2 subscales) than would the UBS Social/Emotional Engagement subscale.

Given the apparent focus of the UBS Social/Emotional Engagement subscale on social and emotional behaviors (see description of subscales in Aim 1 results) and the clear overlap in item content between the subscales, I hypothesized this subscale would show the strongest relationship with the BIMAS-2 Social subscale across convergent and criterion-related validity analyses. I also suspected UBS Social/Emotional Engagement would relate strongly and

⁷ For Aims 3 and 4, the hypotheses and subsequent results are described first for UBS Academic Readiness (i.e., Factor 2) followed by UBS Social/Emotional Engagement (i.e., Factor 1). As the reader will see, this order was chosen to highlight the findings related to UBS Academic Readiness (i.e., Factor 2) which were consistently clearer than the results for UBS Social/Emotional Engagement (i.e., Factor 1).

significantly, although perhaps to a lesser degree, with BIMAS-2 Negative Affect and Conduct subscales due to: (a) the UBS subscale having at least one item seeming to overlap with each BIMAS-2 subscale (e.g., *regulates emotions* with Negative Affect and *cooperation with peers* with Conduct) and (b) the known association between both internalizing and disruptive behavior concerns and impairments in social functioning (Eisenberg et al., 1998; McElwain et al., 2002; Rosen et al., 2014; Rubin et al., 2009). I anticipated the UBS Social/Emotional Engagement subscale would show stronger relationships with these subscales than with BIMAS-2 Cognitive/Attention and Academic Functioning subscales. Related to criterion validity patterns, I hypothesized that UBS Social/Emotional Engagement would show similar patterns in classification accuracy indices, with the best performance shown for classifying BIMAS-2 Social risk status, followed by Negative Affect and Conduct subscales relative to other BIMAS-2 subscales. Furthermore, I anticipated UBS Social/Emotional Engagement would evidence better classification power than UBS Academic Readiness for these three BIMAS-2 subscales.

Convergent Validity

Zero-order bivariate Pearson correlations were conducted as an initial examination of convergent validity between the UBS subscales and total scale scores and BIMAS-2 subscale scores at Q2 (see Table 14). As hypothesized, all correlations between UBS Total/subscales and BIMAS-2 subscales were significant at $p < .001$ and in expected directions based on scoring differences between the two measures (i.e., negative correlations between UBS and BIMAS-2 Behavioral Concern Scales [i.e., Conduct, Negative Affect, and Cognitive/Attention] and positive correlations between UBS and BIMAS-2 Adaptive Scales [i.e., Social and Academic Functioning]). As shown in Table 14, effect sizes ranged from medium (weakest $r = -.43$) to large (strongest $r = -.75$) for correlations between the two measures (Cohen, 1988). Overall,

BIMAS-2 Cognitive/Attention and Academic Functioning subscales showed consistently large effect sizes with UBS subscales and total scale scores ($r = .60$ to $-.75$). In contrast, BIMAS-2 Negative Affect and Social subscales showed relatively weaker correlations with all UBS subscales and total scale scores ($r = -.43$ to $.51$), with all correlations in the medium range or right on the threshold for large (Cohen, 1988). Regarding redundancy, two correlations between the UBS Academic Readiness subscale and the Cognitive/Attention and Academic Functioning BIMAS-2 subscales were found to perhaps suggest redundancy at $r > |.70|$ (Kline, 1979). UBS Total correlations with BIMAS-2 Cognitive/Attention and Academic Functioning subscales were also found to be large at the point of potential redundancy ($r > .70$). All other correlations were arguably distinct ($r < .70$; Kline, 1979).

Table 14

Pearson Bivariate Correlations Between Q2 UBS Total/Subscales and Q2 BIMAS-2 Subscales

	Q2 UBS ($n = 213$)			Q2 BIMAS-2 ($n = 205$)				
	1	2	3	4	5	6	7	8
1 UBS Social/Emo Engagement	--							
2 UBS Academic Readiness	.78	--						
3 UBS Total	.95	.94	--					
4 BIMAS-2 Conduct	-.60	-.47	-.57	--				
5 BIMAS-2 Negative Affect	-.50	-.43	-.50	.67	--			
6 BIMAS-2 Cognitive/Attention	-.67	-.75	-.75	.63	.53	--		
7 BIMAS-2 Social	.49	.46	.51	-.34	-.44	-.29	--	
8 BIMAS-2 Academic Functioning	.60	.74	.71	-.45	-.48	-.71	.50	--

Note. Bold print highlights hypothesized convergent indices. All correlations significant at $p < .001$. Social/Emotional Engagement abbreviated as “Social/Emo Engagement.”

Given all correlations were significant with a medium to large effect size, Fisher's z -tests were used to statistically compare the magnitude of correlations between subscales, particularly related to hypotheses (see bolded correlation coefficients in Table 14). These follow-up contrasts were examined using Fisher r -to- z transformation for dependent correlations (i.e., correlations retrieved from the same sample; Meng et al., 1992; see also Fisher, 1921). Two-tailed tests were used for all comparisons, with values greater than $|1.96|$ considered significant. Results of these comparisons are presented in Table 15. Consistent with hypotheses, BIMAS-2 Cognitive/Attention and Academic Functioning were both significantly more correlated with UBS Academic Readiness than with UBS Social/Emotional Engagement.

Regarding Social/Emotional Engagement, all correlations for the three hypothesized BIMAS-2 subscales (i.e., Social, Negative Affect, and Conduct) were stronger with UBS Social/Emotional Engagement than with UBS Academic Readiness (see Table 14). However, according to results of the Fisher's z -tests (see Table 15), this difference in correlation magnitude was only statistically significant for one BIMAS-2 subscale (i.e., Conduct). Interestingly, as seen in correlation matrix (Table 14), the BIMAS-2 Social subscale was found to have a larger correlation with UBS Total than with either of the UBS subscales. This was the only BIMAS-2 subscale with which UBS Total had a stronger correlation than either UBS subscale. However, the statistical significance of this difference was unable to be explored as the UBS subscale score contributes to the UBS Total score.

Table 15*Fisher's z-Tests of Q2 BIMAS-2 Correlations with Each UBS Subscale at Q2*

Q2 BIMAS-2 Subscale	Q2 UBS Subscale	<i>r</i>	<i>z</i>	<i>p</i>
Academic Functioning	Social/Emotional Engagement	.60	-4.27	< .001
	Academic Readiness	.74		
Cognitive/Attention	Social/Emotional Engagement	-.67	2.59	.009
	Academic Readiness	-.75		
Conduct	Social/Emotional Engagement	-.60	-3.38	< .001
	Academic Readiness	-.47		
Negative Affect	Social/Emotional Engagement	-.50	-1.72	.090
	Academic Readiness	-.43		
Social	Social/Emotional Engagement	.49	0.74	.460
	Academic Readiness	.46		

Note. Bolded text indicates the UBS subscale that correlated significantly more strongly with the BIMAS-2 subscale. BIMAS-2 subscales grouped together in table based on hypotheses: BIMAS-2 Academic Functioning and Cognitive/Attention hypothesized to correlate most strongly with UBS Academic Readiness; BIMAS-2 Conduct, Negative Affect, and Social hypothesized to correlate most strongly with Social/Emotional Engagement.

In examining the correlations between UBS and BIMAS-2 subscales in Table 14, another interesting pattern emerged leading to further post hoc exploration. Along with exploring how BIMAS-2 subscales related differentially to each UBS subscale and total scale, I also examined the patterns in the strength of UBS subscale correlations between hypothesized versus non-hypothesized relationships. Specifically, I anticipated UBS Academic Readiness would have stronger correlations with BIMAS-2 Cognitive/Attention and Academic Readiness subscales

than with subscales hypothesized relate more to UBS Social/Emotional Engagement (i.e., Social, Negative Affect, Conduct). The same was anticipated for UBS Social/Emotional Engagement in showing stronger relationships with hypothesized subscales compared to those hypothesized to relate more to UBS Academic Readiness. As seen in the correlation matrix in Table 14, UBS Academic Readiness correlations showed clear differentiation, with correlations around $r = |.75|$ for hypothesized BIMAS-2 subscales (i.e., Cognitive/Attention and Academic Functioning) compared to around $r = |.45|$ for non-hypothesized subscales. However, correlations for UBS Social/Emotional Engagement with BIMAS-2 subscales showed less clear differentiation. Contrary to hypotheses, the strongest correlation was found between UBS Social/Emotional Engagement and BIMAS-2 Cognitive/Attention ($r = -.67$), followed by Academic Functioning ($r = .60$) and Conduct ($r = -.60$).

Fisher's z -tests were used to statistically compare the magnitude of correlations between UBS Social/Emotional Engagement and hypothesized versus non-hypothesized BIMAS-2 subscales (see Table 16). An example of how to read and understand Table 16 is as follows: the first row in Table 16 describes the correlation between UBS Social/Emotional Engagement and BIMAS-2 Social ($r = .49$) compared to the correlation between UBS Social/Emotional Engagement and BIMAS-2 Academic Functioning ($r = .60$). The comparison between these two correlations using Fisher's z -test produces a significant result ($z = -1.98, p = .048$). Moving to the next row, the correlation between UBS Social/Emotional Engagement and BIMAS-2 Social ($r = .49$) is compared against the correlation between UBS Social/Emotional Engagement and BIMAS-2 Cognitive/Attention ($r = -.67$) and is found to be significantly weaker ($z = 11.87, p < .001$).

Results of Fisher's z -tests revealed UBS Social/Emotional Engagement was significantly more correlated with BIMAS-2 Academic Functioning than all three hypothesized BIMAS-2 subscales (i.e., Social, Negative Affect, Conduct). Additionally, UBS Social/Emotional Engagement was also significantly more correlated with BIMAS-2 Cognitive/Attention than with BIMAS-2 Social and Negative Affect subscales, with no significant differences found between UBS Social/Emotional Engagement correlations with BIMAS-2 Conduct and BIMAS-2 Cognitive/Attention.

Table 16

Fisher's z -Tests Comparing UBS Social/Emotional Engagement Correlations with Hypothesized Versus Non-Hypothesized BIMAS-2 Subscales

Hypothesized BIMAS-2 Subscale	Non-Hypothesized BIMAS-2 Subscale	r	z	p
Social ($r = .49$)	Academic Functioning	.60	-1.98	.048
	Cognitive/Attention	-.67	11.87	< .001
Negative Affect ($r = -.50$)	Academic Functioning	.60	-10.26	< .001
	Cognitive/Attention	-.67	3.31	.001
Conduct ($r = -.60$)	Academic Functioning	.60	-11.57	< .001
	Cognitive/Attention	-.67	1.58	.114

Note. All correlations reported are between the BIMAS-2 subscale indicated in the table and UBS Social/Emotional Engagement. The z -scores presented in the table represent the statistical comparison between the correlation in the first column compared to the correlation in the second column (within the same row).

Concurrent Criterion-Related Validity

A series of logistic regression analyses were conducted with the primary aim of examining the classification accuracy of the UBS Total and subscale scores regarding BIMAS-2 subscale risk status at Q2. For each UBS and BIMAS-2 subscale pairing, the UBS Total or subscale was entered as the predictor variable and BIMAS-2 subscale binary risk status was entered as the criterion variable. Given the significant nature of the correlations between all combinations of UBS Total and subscales and BIMAS-2 subscales, I conducted logistic regression analyses for all 15 possible combinations of the two measures (i.e., two UBS subscales and one total score x five BIMAS-2 subscales). Results from all 15 logistic regressions are exhibited in regression and classification tables organized by UBS scale (i.e., subscales and Total) in Appendices G through I (i.e., see Appendix G for UBS Social/Emotional Engagement, Appendix H for UBS Academic Readiness, and Appendix I for UBS Total regressions). However, based on results of the Fisher's z -tests along with general hypotheses about underlying constructs likely shared between the subscales as described previously, a subset of these comparisons was of particular interest (see bolded values in Table 17). Specifically, UBS Academic Readiness and (a) BIMAS-2 Academic Functioning and (b) BIMAS-2 Cognitive/Attention subscales; and UBS Social/Emotional Engagement and (c) BIMAS-2 Conduct, (d) BIMAS-2 Negative Affect, and (e) BIMAS-2 Social subscales.

As described previously, gender was found to have a significant relationship with BIMAS-2 scores such that female students were rated significantly lower on average (i.e., better functioning) than male students on the BIMAS-2 Cognitive/Attention subscale. As such, a series of multiple logistic regressions were conducted including gender as a simultaneous predictor in the model along with the relevant UBS scale (i.e., Social/Emotional Engagement subscale,

Academic Readiness subscale, and UBS Total). Prior to running these logistic regression analyses, the relationship between gender and each UBS subscale/total score was examined for potential issues of multicollinearity using a series of *t*-tests. These analyses found no significant differences in UBS scores for either subscale or the total score based on student gender, suggesting it was appropriate to include both variables as predictors in the logistic regression models. When gender was added to the model as a covariate it was not found to be a significant predictor of BIMAS-2 Cognitive/Attention risk status in any of the three logistic regressions conducted, while UBS subscale/total score and the overall model statistics remained significant. Additionally, the addition of gender into the model did not seem to improve pseudo R^2 coefficients to any meaningful degree compared to simple logistic regression results with only UBS subscale/Total as a predictor, suggesting little additional benefit achieved in prediction by including both gender and UBS subscale/total scale. Lastly, the classification tables produced from multiple logistic regression analyses were completely unchanged from the simple logistic regression results, suggesting no additional benefit in classification accuracy by adding gender as a covariate in the model. As a result of these nonsignificant results, only the simple regression results are described here and presented in Appendices G, H, and I⁸.

Overall, all 15 of the 15 logistic regression analyses were significant at $p < .001$ (see tables in Appendices G, H, and I), which also held after calculating post hoc Bonferroni correction (i.e., p value of .05 divided by 15 comparisons equals .003). Additionally, both overall model evaluation statistics (i.e., likelihood ratio test and score test) were significant at $p < .001$ for all 15 logistic regression analyses. These statistical indicators suggest the addition of a given

⁸ Results of multiple logistic regression analyses for BIMAS-2 Cognitive/Attention risk status in Aim 3 are available upon request. Simple logistic regression results were prioritized due to ease of interpretation.

UBS subscale or total scale into the models as a predictor significantly improved BIMAS-2 risk category prediction. Said another way, all 15 logistic models with a given UBS subscale or total scale included as a single predictor were more effective than their respective null models.

Hosmer-Lemeshow (H-L) test was examined for all 15 regressions to assess the fit of the logistic model against actual outcomes (i.e., scores indicating risk/concern on BIMAS-2 subscale).

Significant H-L tests (i.e., $p < .05$) indicate potentially poor model fit. Out of the 15 logistic regression analyses, only two models were found to have significant H-L tests: (a) UBS

Social/Emotional Engagement subscale x BIMAS-2 Cognitive/Attention, $\chi^2(6) = 13.75, p = .03$,

and (b) UBS Academic Readiness x BIMAS-2 Cognitive/Attention, $\chi^2(7) = 16.95, p = .02$.

The primary examination of the logistic regression models focused on the degree to which predicted probabilities in the model agreed with actual outcomes based on the reference standard (i.e., BIMAS-2 subscale risk status). The results from all 15 logistic regressions can be viewed in the classification tables included in Appendices G, H, and I, as well as in Table 17 below.

Discussion of these results will focus on the subset of interest (see bolded values in Table 17), with particular attention to sensitivity, specificity, positive predictive value, and negative predictive value estimates. As previously described, high sensitivity (i.e., increased true positives) and high negative predictive value (i.e., low false negatives) estimates were of particular importance for evaluating the UBS. This was because missing an *any risk/concern* case is more detrimental than a false positive at an initial gate of screening (Glover & Albers, 2007). In addition, a moderate PPV was preferred over a high value for the current study, as a lower value allows for a greater proportion of false positives, allowing for a more liberal screening procedure.

Table 17

Aim 3 Concurrent Validity Values: Mean Q2 UBS Scores by Q2 BIMAS-2 Risk/Concern Status and Classification Accuracy Estimates

BIMAS-2 Score in Risk/Concern range for Subscale?	No. of Students	UBS Factor 1 Score	UBS Factor 2 Score	UBS Total Score
		<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>
Conduct				
Yes	31	2.92 (0.65)	2.79 (0.70)	2.86 (0.61)
No	174	3.86 (0.53)	3.64 (0.68)	3.76 (0.56)
Summary	205	3.72 (0.64)	3.51 (0.75)	3.62 (0.65)
Sensitivity (true yes)		38.71%	25.81%	35.48%
Specificity (true no)		96.55%	97.70%	95.98%
PPV		66.67%	66.67%	61.11%
NPV		89.84%	88.08%	89.30%
Hit rate (overall accuracy rate)		87.80%	86.83%	86.83%
Change in % correct from null model		2.9%	1.9%	1.9%
Negative Affect				
Yes	50	3.22 (0.73)	3.02 (0.74)	3.13 (0.68)
No	155	3.87 (0.53)	3.67 (0.68)	3.78 (0.56)
Summary	205	3.71 (0.65)	3.51 (0.75)	3.62 (0.65)
Sensitivity (true yes)		28.00%	22.00%	32.00%
Specificity (true no)		94.84%	94.84%	93.55%
PPV		63.64%	57.89%	61.54%
NPV		80.33%	79.03%	81.01%
Hit rate (overall accuracy rate)		78.54%	77.07%	78.54%
Change in % correct from null model		2.9%	1.5%	2.9%
Cognitive/Attention				
Yes	51	3.07 (0.59)	2.73 (0.61)	2.92 (0.55)
No	154	3.93 (0.51)	3.77 (0.60)	3.86 (0.50)
Summary	205	3.72 (0.65)	3.51 (0.75)	3.63 (0.65)
Sensitivity (true yes)		47.06%	60.78%	52.94%
Specificity (true no)		94.16%	94.81%	92.86%
PPV		72.73%	79.49%	71.05%
NPV		84.30%	87.95%	85.63%
Hit rate (overall accuracy rate)		82.44%	86.34%	82.93%
Change in % correct from null model		7.3%	11.2%	7.8%

Table 17 (Continued)

BIMAS-2 Score in Risk/Concern range for Subscale?	No. of Students	UBS Factor	UBS Factor	UBS Total
		1 Score	2 Score	Score
		<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>
Social				
Yes	40	3.16 (0.74)	3.00 (0.70)	3.09 (0.66)
No	165	3.85 (0.54)	3.63 (0.71)	3.75 (0.58)
Summary	205	3.72 (0.64)	3.51 (0.75)	3.62 (0.64)
Sensitivity (true yes)		35.00%	15.00%	27.50%
Specificity (true no)		97.58%	96.36%	95.76%
PPV		77.78%	50.00%	61.11%
NPV		86.10%	82.38%	84.49%
Hit rate (overall accuracy rate)		85.37%	80.49%	82.44%
Change in % correct from null model		4.9%	0.0%	1.9%
Academic Functioning				
Yes	41	3.02 (0.66)	2.51 (0.57)	2.80 (0.55)
No	164	3.89 (0.52)	3.76 (0.56)	3.83 (0.49)
Summary	205	3.72 (0.65)	3.52 (0.75)	3.63 (0.65)
Sensitivity (true yes)		34.15%	73.17%	56.10%
Specificity (true no)		95.12%	94.51%	95.12%
PPV		63.64%	76.92%	74.19%
NPV		85.25%	93.37%	89.66%
Hit rate (overall accuracy rate)		82.93%	90.24%	87.32%
Change in % correct from null model		2.9%	10.2%	7.3%

Note. Bold print highlights hypothesized convergent indices. Data for the number of students in risk/concern range for BIMAS-2 subscale includes students with scores falling in the *some risk* or *high risk* ranges for Behavioral Concern Scales and in the *concern* range for Adaptive Scales.

Across all UBS subscales and total scales, sensitivity indices (i.e., students indicated as having *any risk/concern* status on a BIMAS-2 subscale who were correctly classified by the UBS) ranged from 15.00% (i.e., UBS Academic Readiness and BIMAS-2 Social) to 73.17% (i.e., UBS Academic Readiness and BIMAS-2 Academic Functioning). Negative predictive

values (i.e., proportion of students with negative screening results who were classified as *low risk/no concern* on a BIMAS-2 subscale) ranged from 79.03% (i.e., UBS Academic Readiness and BIMAS-2 Negative Affect) to 93.37% (i.e., UBS Academic Readiness and BIMAS-2 Academic Functioning). Regarding positive predictive value (i.e., proportion of students with positive screening results who were classified as *any risk/concern* on a BIMAS-2 subscale), values ranged from 50.00% (i.e., UBS Academic Readiness and BIMAS-2 Social) to 79.49% (i.e., UBS Academic Readiness and BIMAS-2 Cognitive/Attention). Specificity values for all UBS and BIMAS-2 subscale examinations were high, ranging from 92.86% (UBS Total and BIMAS-2 Cognitive/Attention) to 97.70% (UBS Academic Readiness and BIMAS-2 Conduct), indicating UBS scales accurately identified the majority of students as true negatives (i.e., students scoring in the *low risk/no concern* range on a given BIMAS-2 subscale).

Overall, the UBS Academic Readiness subscale was found to have the highest classification accuracy indices with BIMAS-2 Cognitive/Attention and Academic Functioning subscales as the reference standard. In evaluating the classification accuracy of the UBS Academic Readiness subscale against the BIMAS-2 Cognitive/Attention, the sensitivity index indicated 60.78% of students identified by BIMAS-2 Cognitive Attention scores in the *any risk* range ($n = 51$) were also identified by the UBS Academic Readiness subscale ($n = 24$). The associated positive predictive power and negative predictive power for these cognitive/attention risk predictions were 79.49% and 87.95%, respectively. Evaluating UBS Academic Readiness against BIMAS-2 Academic Functioning produced a sensitivity index of .732, indicating 73.2% of the students in the *any risk/concern* range on the BIMAS-2 were accurately identified using the UBS. The associated PPV and NPV values for these academic functioning predictions were 77.78% and 86.10%, respectively.

Based on hypotheses, the comparisons of greatest interest for the UBS Social/Emotional Engagement subscale were with: BIMAS-2 (a) Conduct, (b) Negative Affect, and (c) Social. Overall, hypotheses were supported, with comparisons yielding more preferred sensitivity and NPV profiles for BIMAS-2 Conduct (sensitivity = 38.71%, NPV = 89.84%) and Social (sensitivity = 35.00%, NPV = 86.10%) subscales than were found for the UBS Academic Readiness subscale and UBS Total. However, UBS Total was found to produce more preferred sensitivity and NPV indices for identifying students at-risk on the BIMAS-2 Negative Affect subscale (sensitivity = 32.00%, NPV = 81.01%) than were found for UBS Social/Emotional Engagement (sensitivity = 28.00%, NPV = 80.33%) or UBS Academic Readiness (sensitivity = 22.00%, NPV = 79.03%).

In terms of overall classification accuracy, the largest improvements from the null model were found for BIMAS-2 Cognitive/Attention (11.2% increase in correct classification percentage) and BIMAS-2 Academic Functioning (10.2% increase in correct classification percentage) using the UBS Academic Readiness as the predictor. For BIMAS-2 Cognitive/Attention, the addition of UBS Social/Emotional Engagement and UBS Total resulted in a similar increase in percent correctly classified (i.e., 7.3% and 7.8%, respectively). The inclusion of UBS Total as a predictor for BIMAS-2 Academic Functioning risk status saw a change in percentage correct of 7.3% when compared to the null model.

Aim 4: UBS Short-Term Predictive Criterion-Related Validity

Predictive validity of the UBS Total and subscale scores were examined at Q1 using Q4 BIMAS-2 subscales as the criterion measure. This comparison provided a time interval of approximately 7.5 months between administrations. Similar to Aim 3, hypotheses were made for Aim 4 based on results of Aims 1-3. Overall, it was anticipated similar results would be found

for Aim 4 correlational and logistic regression analyses as concurrent analyses in Aim 3, but that the strength of the relationships would be attenuated due to the 7-month time interval between UBS and BIMAS-2 administrations. Specifically, stronger correlations and classification accuracy estimates were predicted for UBS Academic Readiness and BIMAS-2 Academic Functioning and Cognitive/Attention subscales than for other combinations explored. Regarding UBS Social/Emotional Engagement, I anticipated this subscale to have stronger correlations with BIMAS-2 Conduct, Negative Affect, and Social subscales compared to UBS Academic Readiness. Overall, I speculated that all correlation coefficients and logistic regression analyses would be significant and in expected directions, in line with results of Aim 3.

Short-Term Predictive Validity: Correlations

First, predictive validity patterns were explored using zero-order bivariate Pearson correlations between UBS subscales and total scores at Q1 and BIMAS-2 subscale scores at Q4 (see Table 18). As hypothesized, all correlations between Q1 UBS and Q4 BIMAS-2 subscales and/or total scale were significant at $p < .001$ and in expected directions based on scoring differences between the two measures (i.e., negative correlations between UBS and BIMAS-2 Behavioral Concern scales and positive correlations between UBS and BIMAS-2 Adaptive scales). Effect sizes for correlations between Q1 UBS and Q4 BIMAS-2 scales ranged from medium (weakest $r = -.35$) to large (strongest $r = -.61$; Cohen, 1988). Only three correlations showed large effect sizes: Q1 UBS Academic Readiness and Q4 BIMAS-2 Cognitive/Attention ($r = -.61$), Q1 UBS Academic Readiness and Q4 BIMAS-2 Academic Functioning ($r = .56$ subscales), and Q1 UBS Total and Q4 BIMAS-2 Cognitive/Attention ($r = -.56$). Regarding redundancy between the two measures, no correlation coefficients were greater than $|.70|$ suggesting all correlations were distinct ($r < .70$; Kline, 1979).

Table 18*Pearson Bivariate Correlations Between Q1 UBS Total/Subscales and Q4 BIMAS-2 Subscales*

	Q1 UBS (<i>n</i> = 214)			Q4 BIMAS-2 (<i>n</i> = 211)				
	1	2	3	4	5	6	7	8
1 Q1 UBS Social/Emo Engagement	--							
2 Q1 UBS Academic Readiness	.81	--						
3 Q1 UBS Total	.96	.94	--					
4 Q4 BIMAS-2 Conduct	-.44	-.39	-.44	--				
5 Q4 BIMAS-2 Negative Affect	-.35	-.42	-.40	.56	--			
6 Q4 BIMAS-2 Cognitive/Attention	-.47	-.61	-.56	.63	.57	--		
7 Q4 BIMAS-2 Social	.36	.40	.40	-.34	-.36	-.32	--	
8 Q4 BIMAS-2 Academic Functioning	.37	.56	.48	-.43	-.56	-.75	.50	--

Note. All significance levels at $p < .001$. $n = 198$ for all correlations between Q1 UBS and Q4 BIMAS-2. Bold print highlights hypothesized convergent indices. Social/Emotional Engagement abbreviated as “Social/Emo Engagement.”

Given all correlations were significant with a medium to large effect size, Fisher’s z -tests were used to statistically compare the magnitude of correlations between subscales, particularly related to hypotheses (see bolded correlation coefficients in Table 18). These follow-up contrasts were examined using Fisher r -to- z transformation for dependent correlations (i.e., correlations retrieved from the same sample; Meng et al., 1992; see also Fisher, 1921). Two-tailed tests were used for all comparisons, with values greater than $|1.96|$ considered significant. Results of these statistical comparisons are presented in Table 19. Consistent with predictions and findings for Aim 3, BIMAS-2 Cognitive/Attention and Academic Functioning were both significantly more correlated with UBS Academic Readiness than with UBS Social/Emotional Engagement.

Table 19*Fisher's z-Tests for Q4 BIMAS-2 Correlations with Each Q1 UBS Subscale*

Q4 BIMAS-2 Subscale	Q1 UBS Subscale	<i>r</i>	<i>z</i>	<i>p</i>
Academic Functioning	Social/Emotional Engagement	.37	-4.94	< .001
	Academic Readiness	.56		
Cognitive/Attention	Social/Emotional Engagement	-.47	3.86	< .001
	Academic Readiness	-.61		
Conduct	Social/Emotional Engagement	-.44	-1.26	.209
	Academic Readiness	-.39		
Negative Affect	Social/Emotional Engagement	-.35	1.73	.083
	Academic Readiness	-.42		
Social	Social/Emotional Engagement	.36	-0.99	.313
	Academic Readiness	.40		

Note. Bolded text indicates the UBS subscale that correlated significantly more strongly with the BIMAS-2 subscale. BIMAS-2 subscales grouped together in table based on hypotheses: Academic Functioning and Cognitive/Attention hypothesized to correlate most strongly with UBS Academic Readiness; Conduct, Negative Affect, and Social hypothesized to correlate most strongly with Social/Emotional Engagement.

I predicted Q1 UBS Social/Emotional Engagement would have a stronger relationship with Q4 BIMAS-2 (a) Conduct, (b) Negative Affect, and (c) Social subscales, than Q1 UBS Academic Readiness with those Q4 BIMAS-2 subscales, in a similar manner as was found for Aim 3. Results of correlation analyses were not consistent with this hypothesis. Looking at the correlation matrix in Table 18, the only hypothesized BIMAS-2 subscale that showed a stronger correlation with UBS Social/Emotional Engagement than with UBS Academic Readiness was

BIMAS-2 Conduct. The other two hypothesized BIMAS-2 subscales (i.e., Social and Negative Affect) were found to have stronger correlations with UBS Academic Readiness, contrary to expectations. However, results of Fisher's z -tests (see Table 19) found none of these differences in correlations between the two UBS subscales as statistically significant. Specifically, there was no significant difference in the magnitude of correlations between UBS Social/Emotional Engagement and hypothesized BIMAS-2 subscales (i.e., Social, Negative Affect, Conduct) and the correlations between UBS Academic Readiness and those same three BIMAS-2 subscales.

Short-Term Predictive Validity: Logistic Regressions

A series of logistic regression analyses were conducted to evaluate short-term predictive validity of the UBS at Q1 with regards to BIMAS-2 risk classification at Q4. For each UBS Total/subscale and BIMAS-2 subscale pairing, except for all UBS pairings with BIMAS-2 Cognitive/Attention, UBS Total or subscale was entered as the sole predictor variable and BIMAS-2 subscale binary risk status was entered as the criterion variable. As in Aim 3, logistic regression analyses allowed for the examination of the classification accuracy of the UBS. Given the significant nature of the correlations between all combinations of Q1 UBS Total and subscales and Q4 BIMAS-2 subscales, I conducted logistic regression analyses for all 15 possible combinations of the two measures (i.e., three UBS scales and five BIMAS-2 scales). Results from all 15 logistic regressions are exhibited in regression and classification tables organized by UBS Total/subscale in Appendices J through L (i.e., see Appendix J for UBS Social/Emotional Engagement, Appendix K for UBS Academic Readiness, and Appendix L for UBS Total regressions). However, as with Aim 3, a subset of these comparisons was of particular interest based on results of Fisher's z -tests and hypothesized relationships (see bolded text in Table 20). Specifically, UBS Academic Readiness and (a) BIMAS-2 Academic

Functioning and (b) BIMAS-2 Cognitive/Attention subscales; and UBS Social/Emotional Engagement and (a) BIMAS-2 Conduct, (b) BIMAS-2 Negative Affect, and (c) BIMAS-2 Social subscales.

Similar to results in Aim 3, gender was found to have a significant relationship with Q4 BIMAS-2 Cognitive/Attention *T*-scores such that female students were rated significantly lower (i.e., better functioning) on average than male students. Based on this finding, an additional series of multiple logistic regression analyses were also run to include gender as a simultaneous predictor in the logistic regressions along with the relevant UBS scale (i.e., Social/Emotional Engagement subscale, Academic Readiness subscale, and UBS Total) at Q1 predicting BIMAS-2 Cognitive/Attention risk status at Q4. The results of these two-predictor models can be found in Appendices J (Tables J5 and J6), K (Tables K5 and K6), and L (Tables L5 and L6). Like Aim 3, the relationship between each UBS subscale/total score and gender was examined for potential multicollinearity through a series of *t*-tests. Gender was not found to significantly influence UBS scores and thus, both predictors were deemed appropriate to include simultaneously in the models. For all three logistic regressions predicting BIMAS-2 Cognitive/Attention risk status, both UBS score (i.e., Total or subscale; predictor) and gender (i.e., covariate) were found to be significant predictors in the model. See Tables J5, K5, and L5 in Appendices for multiple logistic regression results and interpretation. Overall, results from all three multiple logistic regressions indicated female students were less likely than male students to be rated by teachers in the *any risk* (i.e., *some* or *high risk*) range on the BIMAS-2 Cognitive/Attention subscale (holding UBS score constant). UBS subscale and total scale score remained a significant predictor, with each one-point increase in a student's UBS score resulting in decreased odds of that student scoring in the *any risk* range on the BIMAS-2 Cognitive/Attention subscale (holding gender constant). The

specific interpretations of these results are included in the Tables J5, K5, and L5. In terms of the primary metrics of interest (i.e., classification accuracy), the sensitivity values for UBS Total and subscale scores predicting BIMAS-2 subscale risk status were modestly improved by including gender into the model. Specifically, when gender was included in the model, the sensitivity index for UBS Social/Emotional Engagement predicting BIMAS-2 Cognitive/Attention risk status improved from 21.3% to 23.4%; however, both specificity and overall accuracy decreased slightly with the addition of gender in the model (i.e., decrease of 3.4 percentage points for specificity and 2.0 percentage points for overall accuracy). Similarly, for UBS Academic Readiness, sensitivity increased from 44.7% to 48.9% when gender was included in the model, whereas specificity remained the same and overall accuracy increased by 1.0 percentage points. For UBS Total, sensitivity increased from 34.0% to 40.4%, specificity improved from 93.4% to 94.0%, and overall accuracy improved from 79.3% to 81.3%. The results of the three multiple logistic regression analyses related to classification accuracy are presented in Tables J6, K6, and L6 in Appendices J, K, and L, as well as in Table 20.

Overall, 15 out of 15 logistic regression analyses were significant at $p < .001$ (see Appendices J, K, and L). Additionally, both overall model evaluation statistics (i.e., likelihood ratio test and score test) were significant at $p < .001$ for all 15 logistic regression analyses. These statistical indicators suggest all 15 logistic models with UBS Total or subscale included as a single predictor were more effective than their respective null models. Hosmer-Lemeshow (H-L) test was examined for all 15 regressions to assess the fit of the logistic model against actual outcomes (i.e., scores indicating risk/concern on BIMAS-2 subscale). Significant H-L tests (i.e., $p < .05$) indicate potentially poor model fit. Out of the 15 logistic regression analyses, five models (i.e., three single-predictor logistic regression models and two multiple predictor logistic

regression models) were found to have significant H-L tests: (a) Q1 UBS Social/Emotional Engagement subscale predicting Q4 BIMAS-2 Conduct risk status, $\chi^2(6) = 14.43$, $p = .025$; (b) Q1 UBS Academic Readiness predicting Q4 BIMAS-2 Conduct risk status, $\chi^2(6) = 15.54$, $p = .016$; (c) Q1 UBS Total predicting Q4 BIMAS-2 Conduct risk status, $\chi^2(7) = 19.19$, $p = .008$; (d) Q1 UBS Social/Emotional Engagement subscale and gender predicting Q4 BIMAS-2 Cognitive/Attention risk status, $\chi^2(8) = 15.85$, $p = .045$; and (e) Q1 UBS Total and gender predicting Q4 BIMAS-2 Cognitive/Attention risk status, $\chi^2(8) = 17.806$, $p = .023$.

The primary examination of the logistic regression models focused on the degree to which predicted probabilities in the model agreed with actual outcomes based on the reference standard (i.e., BIMAS-2 subscale risk status). The full results from all 15 logistic regressions can be viewed in the classification tables included in Appendices J, K, and L. Discussion of these results will focus on the subset of interest (see bolded text in Table 20), with particular attention to sensitivity, specificity, positive predictive value, and negative predictive value estimates (see Table 20). The goals for Aim 4 mirror that of Aim 3, such that high sensitivity, high NPV, and moderate PPV were preferred to prioritize over-identification of students at risk of potential social, emotional, and behavioral challenges as opposed to under-identify.

Table 20

Aim 4 Predictive Validity Values: Mean Q1 UBS Scores by Q4 BIMAS-2 Risk/Concern Status and Classification Accuracy Estimates

BIMAS-2 Score in Risk/Concern range for Subscale?	No. of Students	UBS Factor 1 Score	UBS Factor 2 Score	UBS Total Score
		<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>
Conduct				
Yes	27	2.88 (0.95)	2.65 (0.93)	2.78 (0.90)
No	171	3.65 (0.60)	3.35 (0.71)	3.52 (0.61)
Summary	198	3.55 (0.71)	3.25 (0.78)	3.42 (0.70)
Sensitivity (true yes)		18.52%	7.41%	14.81%
Specificity (true no)		98.83%	99.42%	99.42%
PPV		71.43%	66.67%	80.00%
NPV		88.48%	87.18%	88.08%
Hit rate (overall accuracy rate)		87.88%	86.87%	87.88%
Change in % correct from null model		1.5%	0.5%	1.5%
Negative Affect				
Yes	33	2.88 (0.92)	2.61 (0.82)	2.76 (0.84)
No	165	3.68 (0.57)	3.38 (0.71)	3.55 (0.59)
Summary	198	3.55 (0.71)	3.25 (0.78)	3.42 (0.70)
Sensitivity (true yes)		21.21%	18.18%	21.21%
Specificity (true no)		98.18%	98.18%	100.00%
PPV		70.00%	66.67%	100.00%
NPV		86.17%	85.71%	86.39%
Hit rate (overall accuracy rate)		85.35%	84.85%	86.87%
Change in % correct from null model		2.1%	1.5%	3.6%
Cognitive/Attention				
Yes	47	3.04 (0.85)	2.57 (0.84)	2.83 (0.80)
No	151	3.70 (0.57)	3.47 (0.63)	3.60 (0.56)
Summary	198	3.54 (0.70)	3.26 (0.78)	3.42 (0.70)
Sensitivity (true yes)		23.40%	48.94%	40.43%
Specificity (true no)		94.04%	93.38%	94.04%
PPV		55.00%	69.70%	67.86%
NPV		79.78%	85.45%	83.53%
Hit rate (overall accuracy rate)		77.27%	82.83%	81.31%
Change in % correct from null model		1.0%	6.5%	5.0%

Table 20 (Continued)

BIMAS-2 Score in Risk/Concern range for Subscale?	No. of Students	UBS Factor 1 Score <i>M (SD)</i>	UBS Factor 2 Score <i>M (SD)</i>	UBS Total Score <i>M (SD)</i>
Social				
Yes	17	2.92 (0.92)	2.49 (0.80)	2.73 (0.84)
No	181	3.60 (0.65)	3.33 (0.74)	3.48 (0.65)
Summary	198	3.54 (0.70)	3.26 (0.78)	3.42 (0.70)
Sensitivity (true yes)		5.88%	5.88%	5.88%
Specificity (true no)		98.34%	99.45%	98.34%
PPV		25.00%	50.00%	33.33%
NPV		91.75%	91.84%	91.75%
Hit rate (overall accuracy rate)		90.40%	91.41%	90.40%
Change in % correct from null model		-1.0%	0.0%	-1.0%
Academic Functioning				
Yes	35	3.01 (0.85)	2.46 (0.70)	2.76 (0.76)
No	163	3.66 (0.61)	3.43 (0.69)	3.56 (0.61)
Summary	198	3.55 (0.70)	3.26 (0.78)	3.42 (0.71)
Sensitivity (true yes)		14.29%	31.43%	25.71%
Specificity (true no)		98.77%	95.09%	98.16%
PPV		71.43%	57.89%	75.00%
NPV		84.29%	86.59%	86.02%
Hit rate (overall accuracy rate)		83.84%	83.84%	85.35%
Change in % correct from null model		1.5%	1.5%	3.1%

Note. Data for the number of students in *any risk/concern* range for BIMAS-2 subscale includes students with scores falling in the *some risk* or *high risk* ranges for Behavioral Concern Scales and in the *concern* range for Adaptive Scales.

Across both UBS subscales and total scales, sensitivity indices ranged from 5.88% (for all UBS subscales/Total scale and BIMAS-2 Social) to 48.94% (UBS Academic Readiness with gender as a covariate and BIMAS-2 Cognitive/Attention). Negative predictive values ranged from 79.78% (UBS Social/Emotional Engagement with gender as a covariate and BIMAS-2

Cognitive/Attention) to 91.84% (UBS Academic Readiness and BIMAS-2 Social). Regarding positive predictive value, values ranged from 25.00% (UBS Social/Emotional Engagement and BIMAS-2 Social) to 100.00% (UBS Total and BIMAS-2 Negative Affect). Specificity values for all UBS and BIMAS-2 examinations were high, ranging from 93.38% (UBS Academic Readiness with gender as a covariate and BIMAS-2 Cognitive/Attention) to 100.00% (UBS Total and BIMAS-2 Negative Affect), indicating all UBS Total and subscales accurately identified most students as true negatives in terms of risk status across all BIMAS-2 subscales (i.e., students scoring in the *low risk/no concern* range on a given BIMAS-2 subscale).

Related to hypotheses, UBS Academic Readiness subscale showed the most preferable combinations of classification accuracy indices, with the highest values for sensitivity compared to UBS Total and Social/Emotional Engagement (i.e., 48.94% for BIMAS-2 Cognitive/Attention with gender included as a covariate; 31.43% for BIMAS-2 Academic Functioning), moderate values for PPV (i.e., 69.70% for Cognitive/Attention with gender included as a covariate; 57.89% for Academic Functioning); and higher values for NPV compared to UBS Total and Social/Emotional Engagement on both BIMAS-2 Cognitive/Attention (i.e., 85.45% with gender included as a covariate) and Academic Functioning (86.59%) subscales. These sensitivity indices indicate 48.94% of students identified by BIMAS-2 Cognitive/Attention scores as being in the *any risk* range ($n = 47$) and 31.43% of students identified by BIMAS-2 Academic Functioning scores as being in the *concern* range ($n = 35$) were predicted by UBS Academic Readiness subscale scores as falling in the *any risk/concern* range.

Based on hypotheses, the comparisons of greatest interest for the UBS Social/Emotional Engagement subscale were with: BIMAS-2 (a) Conduct, (b) Negative Affect, and (c) Social. Overall, UBS Total and both subscales, including UBS Social/Emotional Engagement, had poor

classification accuracy indices for predicting BIMAS-2 Social subscale risk status. Specifically, sensitivity values were 5.88% for UBS Total and both subscales, and overall classification accuracy either did not improve from null model (i.e., UBS Academic Readiness) or worsened (i.e., for UBS Social/Emotional Engagement and Total) with the addition of UBS Total or subscale as predictor. While UBS Social/Emotional Engagement showed the highest sensitivity values for BIMAS-2 Conduct (18.52%) and Negative Affect (21.21%) compared to UBS Total and Academic Readiness, these indicate that only 18.52% and 21.21% of students identified in the *any risk* range on BIMAS-2 Conduct and Negative Affect subscales, respectively, were also identified as *any risk* by UBS Social/Emotional Engagement subscale scores. NPV values for UBS Social/Emotional Engagement as the predictor were relatively high for both BIMAS-2 Conduct (88.48%) and Negative Affect (86.17%) subscales.

In terms of overall classification accuracy, the largest improvements from the null model were found for BIMAS-2 Cognitive/Attention (increase of 6.5% percentage points in correct classification percentage) with UBS Academic Readiness and gender as predictors and for BIMAS-2 Cognitive/Attention with UBS Total and gender as the predictors (increase of 5.0% percentage points in correct classification percentage) with UBS Total and gender as predictors. The greatest increase with UBS as a single predictor was found for Negative Affect (increase of 3.6% percentage points in correct classification percentage) with UBS Total as the predictor. The remaining UBS and BIMAS-2 total/subscale combinations resulted in minimal change in percentage correct from the null model, with the majority falling between -1.0 (i.e., worse predictive accuracy) to 1.5 in terms of change in accuracy percentage due to the included predictor(s).

Discussion

The current study was a preliminary examination of the psychometric properties of the locally developed UBS in a sample of students enrolled in a public elementary school in Hawai‘i. The UBS is a brief, nine-item universal screening tool intended to align with core SEL competencies and GLOs prioritized by the HDOE, meet the school’s needs for a free and less burdensome measure than the BIMAS-2, and designed to identify youth at risk of social, emotional, and/or behavioral difficulties.

The goal of the current study was to examine the factor structure of the UBS and assess preliminary evidence for its reliability and validity across subscales and total scale scores. Consistent with initial speculations, results of this study supported a two-factor structure for the UBS with one factor related to social and emotional competencies and the second factor related to learning behaviors and classroom-specific competencies. Regarding reliability, study findings suggested adequate internal consistency and test-retest reliability of UBS Total scale and subscale scores with the current sample. Results were mixed regarding convergent validity and concurrent criterion-related validity of the UBS Total and subscale scores when using BIMAS-2 subscale *T*-scores and risk status, respectively, as the criteria. Specifically, convergent validity patterns related to correlations between UBS and BIMAS-2 subscales mostly emerged in line with my speculations. Unexpectedly, however, *both* UBS subscales related most strongly to BIMAS-2 Cognitive/Attention and Academic Functioning subscales, not just UBS Academic Readiness as anticipated. Concurrent classification accuracy estimates were more promising for hypothesized relationships between UBS and BIMAS-2 subscales than non-hypothesized relationships. Overall, UBS subscales were found to have high specificity and NPV values, but overall poor sensitivity values for classifying current student risk status as indicated by BIMAS-

2 *T*-scores. Results of predictive criterion validity analyses were less consistent with hypotheses overall. Specifically, UBS Academic Readiness evidenced stronger correlations with all BIMAS-2 subscales (except Conduct) than those found for UBS Social/Emotional Engagement, including Social and Negative Affect subscales, which were hypothesized to relate more strongly to UBS Social/Emotional Engagement. Furthermore, all logistic regression models significantly predicted future BIMAS-2 subscale risk status over a 7.5-month interval regardless of UBS Total scale or subscale included as the predictor. However, although specificity and NPV estimates were high for all UBS and BIMAS-2 subscale comparisons, sensitivity values were considerably worse for predicting students' end of year BIMAS-2 subscale risk status compared to those found for concurrent validity.

Major Findings

Aim 1: UBS Factor Structure

Regarding the study's first aim, an exploratory factor analysis of the UBS produced two factors representing the constructs of (a) Social/Emotional Engagement and (b) Academic Readiness. This two-factor structure aligned with my initial speculation that items might separate into two categories: items focused on social and emotional aspects of functioning and items focused on academic or learning behaviors. The items loading onto the Social/Emotional Engagement subscale included all of those corresponding with CASEL's (2015) core SEL competencies as well as the two socially focused GLOs (i.e., community contributor and effective communicator). Items loading onto the Academic Readiness subscale correspond directly with four of the six GLO competencies (i.e., self-directed learner, complex thinker, quality producer, and effective and ethical user of technology) and did not appear to directly correspond with CASEL's (2015) SEL competency framework. A sufficient number of items

(i.e., greater than three) loaded clearly and significantly onto each factor, suggesting adequate item composition for each subscale. Furthermore, the identified factor structure held regardless of EFA methodology used and was replicated for each administration event, suggesting additional confidence in the UBS factor structure identified in this study and stability of the factors within this sample.

The benefit of the UBS two-factor structure over the single factor structure appears to also be supported by results of criterion validity aims. As was revealed in the results for Aims 3 and 4 of the current study, UBS subscales tended to outperform UBS Total scale scores in predicting both current and future BIMAS-2 risk status. Thus, subscales appear to provide additional differentiation that might be beneficial in maximizing measure performance compared to using only UBS Total score. As such, utilization of the full UBS measure seems warranted at this time and schools may gain the most benefit from the measure by examining all three UBS scores for students separately (e.g., both subscale scores and total score). Lastly, students' scores on the UBS Social/Emotional Engagement and UBS Academic Readiness subscales were strongly correlated ($r = .78$), which aligns with my initial expectations based on the SEL literature noting the impact of SEL interventions (i.e., improving students' social and emotional skills) on academic outcomes such as grade point average and standardized test scores (CASEL, 2003; SAMHSA, 2002; Corcoran et al., 2018; Durlak et al., 2011).

Aim 2: UBS Reliability

Overall, findings across analyses appear to suggest preliminary support for the reliability of UBS Total and subscale scores in the current sample. Cronbach's alpha coefficients of UBS Total and both subscale scores exceeded Nunnally and Bernstein's (1994) $\alpha \geq .80$ threshold for internal consistency for all four quarters of administration. In addition, Cronbach's alpha

coefficients for UBS subscales and Total scale were more favorable than those found for BIMAS-2 subscales with the current sample, except for Cognitive/Attention which was found to be comparable. Consistent with tentative hypotheses, UBS Total scale and both subscales demonstrated initial evidence supporting stability of scores over approximately 2-, 3-, 5-, and 7-month intervals within a single academic year, with all correlations in the excellent range (Cohen, 1988). Additionally, larger correlations were found between adjacent UBS administrations with correlations decreasing as time intervals increased (i.e., Q1 to Q2 v. Q1 to Q3), which is consistent with other examinations of screening measures (Epstein & Sharma, 1998; Lane, Menzies, et al., 2012). Additionally, aside from the relatively low test-retest coefficient found for the BIMAS-2 Social subscale (i.e., $r = .50$), UBS subscale and Total scale ($r = .71$ to $.75$) and BIMAS-2 subscale correlation coefficients ($r = .63$ to $.79$) were comparable for the 5-month time interval between Q2 and Q4. Overall, these findings provide initial evidence to support the internal consistency reliability of UBS subscale and total scale scores, as well as the stability of these data within the current sample. Furthermore, these reliability estimates appear consistent with other published screening measures with a similar focus on SEL competencies.

Aim 3: UBS Convergent and Concurrent Criterion-Related Validity

Regarding my third aim⁹, a series of correlations and logistic regression analyses were performed to examine the convergent validity and concurrent criterion-related validity of UBS subscales and total scale scores using BIMAS-2 subscale risk-status as the criterion. Overall, all

⁹ It is important to note a necessary caveat to the discussion of results for Aims 3 and 4. All results regarding UBS validity are interpreted within the context of the BIMAS-2 as the criterion measure, which is a significant limitation of the current study (see limitations and future directions section for detailed description). As such, the reader is cautioned to avoid overgeneralizing results of these aims past the comparison of UBS performance relative to that of the BIMAS-2.

correlations between the three UBS scales (i.e., two subscales and total scale) and the five BIMAS-2 subscales were significant and in expected directions (i.e., positive for Adaptive Scales, negative for Behavioral Concern scales) with medium to large effect sizes. Results of simple logistic regression analyses¹⁰ provided initial support for concurrent validity between the two measures as all models were significant, and all goodness-of-fit indices suggested adequate fit for 13 out of the 15 concurrent validity models. For the two models with significant Hosmer-Lemeshow (H-L) goodness-fit-tests, however, there was insufficient evidence to suggest substantial concerns with model interpretability given all other indices suggested adequate fit. Furthermore, the H-L test is sensitive to large sample sizes, and potentially irrelevant discrepancies may lead to rejection of the hypothesis of perfect fit (Nattino et al., 2020). Thus, perhaps these two models are also adequately fitted and interpretable. Taken together, these findings suggested prediction of BIMAS-2 subscale risk status was significantly improved by including any UBS scale (i.e., total scale or subscales) as a predictor regardless of the BIMAS-2 subscale¹¹.

Digging deeper into the overall results of this study aim, I will first explore findings related to the convergent validity of each UBS subscale with a particular focus on hypothesized versus non-hypothesized relationships, then, I will discuss results related to the concurrent validity of each subscale with attention to relevant implications. Although all combinations of

¹⁰ Although results of ANOVAs found teachers, on average, rated male students as having more problems than female students on the Cognitive/Attention subscale, gender was not a significant predictor of students' Q2 Cognitive/Attention risk status when included as a covariate in two-predictor logistic regression models.

¹¹ It is important to note this set of analyses involved a high number of comparisons using the same data set and variables and thus, to correct for potential increase in Type I error rate, Bonferroni corrections were calculated post hoc for the number of logistic regression analyses (i.e., $n = 15$) to compute a corrected p -value (i.e., Corrected p -value = $.001 [p\text{-value}] * 15$ [number of tests]; Bonferroni, 1936). This resulted in a corrected $p = .015$ threshold, which still held up for results in this part of my study. Thus, these significant logistic regression analyses still carry the same 5% chance of false positive result and can be interpreted as likely representative of the true relationship underlying these variables within the current study.

UBS and BIMAS-2 subscale correlations were significant, there were varying levels of differentiation between the two UBS subscales. As predicted, BIMAS-2 Cognitive/Attention and Academic Functioning subscales evidenced significantly stronger correlations with UBS Academic Readiness than UBS Social/Emotional Engagement. The magnitude of the correlations between these BIMAS-2 subscales and UBS Academic Readiness were also strong enough to perhaps indicate redundancy, potentially suggesting that limited additional information about student functioning in these domains are gained from administering both measures concurrently compared to only one of them. Additionally, in comparing the absolute values of the correlations between UBS Academic Readiness and each of the five BIMAS-2 subscales, correlations between UBS Academic Readiness and hypothesized scales (i.e., Cognitive/Attention and Academic Functioning) were significantly stronger than those found between UBS Academic Readiness and non-hypothesized subscales (i.e., Conduct, Negative Affect, and Social). These findings may suggest tentative evidence for the construct validity of the UBS Academic Readiness subscale scores, such that this subscale showed more convergence with the subscales understood to measure similar constructs and slight divergence (i.e., weaker but still significant correlations) with subscales understood to measure less related constructs. The use of the word *tentative* is important, however, as this interpretation is focused on significant differences between UBS and BIMAS-2 correlations that were all found to be significant at $p < .001$ and thus, likely does not represent a sufficient degree of evidence for discriminant validity. Overall, however, as compared to the UBS Social/Emotional Engagement subscale discussed below, the UBS Academic Readiness subscale showed the greatest degree of convergent validity and potential discrimination regarding association with BIMAS-2 subscales. It is important to note these results may be influenced by potential measurement or item

contamination effects. While none of the items on the UBS were taken directly from the BIMAS-2, UBS items on the Academic Readiness subscale appear to be very similar to at least four of the seven items on BIMAS-2 Cognitive/Attention subscale and three of the five items on the Academic Functioning subscale. Thus, it is possible the strong relationship between these UBS and BIMAS-2 subscales is due to the similarity in item content and thus, does not truly represent convergence between two independent measures.

The UBS Social/Emotional Engagement subscale, on the other hand, did not show the same degree of differentiation with regard to its correlations with BIMAS-2 subscales. As predicted, BIMAS-2 Conduct, Negative Affect, and Social subscales correlated more strongly with UBS Social/Emotional Engagement than UBS Academic Readiness, however this difference was only statistically significant for BIMAS-2 Conduct. Additionally, the magnitude of the absolute values for these correlations did not reach redundancy, suggesting potentially less overlap in constructs underlying UBS Social/Emotional Engagement and these three BIMAS-2 subscales. I hypothesized these BIMAS-2 subscales would relate more to UBS Social/Emotional Engagement than UBS Academic Readiness due to the known associations between social-emotional functioning and disruptive behavior (i.e., Conduct; Rosen et al., 2014; McElwain et al., 2002) and mood and anxiety concerns (Negative Affect; Eisenberg et al., 1998; Rubin et al., 2009). Based on the degree of similarity between items and underlying constructs, I also suspected UBS Social/Emotional Engagement would relate particularly strongly with BIMAS-2 Social. Post hoc Fisher's z transformations revealed that contrary to my speculations, UBS Social/Emotional Engagement evidenced significantly stronger correlations with BIMAS-2 Academic Functioning and Cognitive/Attention (i.e., non-hypothesized) subscales than with BIMAS-2 Negative Affect and Social (i.e., hypothesized) subscales. Furthermore, although still

evidencing a medium effect size, the weakest correlation for UBS Social/Emotional Engagement was found with the BIMAS-2 Social subscale. This finding is particularly noteworthy given the apparent focus and original intention of the UBS Social/Emotional Engagement subscale for assessing social and emotional functioning. It is possible one potential reason for the lack of apparent overlap between these social-focused subscales relates to issues with item construction on the UBS. Specifically, although seeming to assess similar constructs, the descriptions included for each UBS Social/Emotional Engagement subscale item appear to (a) provide multiple different behavioral considerations for rating the items (e.g., Item 2: *Uses socially appropriate responses: Responds appropriately to the emotions/behaviors of others, is an effective communicator*) and (b) use terminology that may be ambiguous and unclear in terms of observable behavior (e.g., community contributor, effective communicator, demonstrates growth mindset). Compared to the BIMAS-2 Social subscale, which asks teachers to rate students on six clear and observable behaviors, there is more opportunity on the UBS for teacher interpretation of the items which could impact the relation between these two subscales. Regarding the lack of relation with BIMAS-2 Negative Affect, only one item on the UBS Social/Emotional Engagement subscale directly references the student's own emotional experience, while others focus on the student's interaction with others' emotions and their social environment. Furthermore, UBS Social/Emotional Engagement items do not focus on behaviors indicative of anxiety or depression, and thus, it may not be surprising they did not relate as strongly with BIMAS-2 Negative Affect subscale. Similarly, of the five items consistently rated on the BIMAS-2 Conduct subscale (i.e., four items exhibited little to no use) two of those items relate to emotional expression (i.e., appeared angry, lost temper) and two relate to social interaction (i.e., was aggressive/bullied others, fought with others). The inclusion of socially-focused

behaviors on this subscale might have influenced the stronger associations found between UBS Social/Emotional Engagement and Conduct compared to Negative Affect. Overall, evidence for the UBS Social/Emotional Engagement subscale's validity was not clearly established through these analyses. Therefore, stakeholders interested in assessing student concerns related to social functioning, conduct, and negative affect are cautioned against relying on UBS Social/Emotional Engagement scale at this time and instead should consider administering other screeners known to be valid and reliable measures of these constructs. More research is needed to fully understand the findings of the current study and whether these results are specific to the current sample and context or are representative of the true relationships between these measures and underlying constructs.

Next, I will discuss the concurrent validity findings for the UBS Academic Readiness and UBS Social/Emotional engagement subscales. While logistic regression analyses were all significant, the main psychometric indices of interest in examining concurrent criterion-related validity were sensitivity, specificity, PPV, and NPV. As previously mentioned, sensitivity was prioritized over specificity, high NPVs were preferred, and moderate (i.e., not high) PPVs, with high sensitivity deemed the most critical of the four indices. Overall accuracy percentages are not included in this discussion of results as these were not informative due to the high proportion of true negatives in the sample overly inflating overall accuracy percentages (Glover & Albers, 2007; Trevethan, 2017). Specifically, the disproportionately large subsample of students identified as not at risk for social, emotional, or behavioral concerns resulted in specificity values having greater weight in the overall accuracy percentage.

All sensitivity values found for UBS subscales and Total scale scores were under the 75% acceptability threshold noted in the literature (Glover & Albers, 2007), including those for the

UBS Academic Readiness subscale. Of all 15 comparisons explored between UBS and BIMAS-2 scales, the only sensitivity values found to be greater than 50% and coming closest to the 75% acceptability threshold were those between the UBS Academic Engagement subscale and hypothesized BIMAS-2 subscales. The UBS Academic Engagement subscale evidenced the largest sensitivity values in predicting concurrent BIMAS-2 Academic Functioning risk status, with sensitivity (73.17%) and specificity (94.51%) values suggesting a 26.83% false negative error rate and a 5.49% false positive error rate, respectively. The next highest sensitivity value was found for UBS Academic Engagement subscale predicting BIMAS-2 Cognitive/Attention, with sensitivity (60.78%) and specificity (94.81%) values suggesting a 39.22% false negative error rate and a 5.19% false positive error rate. The NPVs for these sets of comparisons were each above the more conservative .80 (i.e., 80%) benchmark reported in the literature (see Glover & Albers, 2007). While these findings supported hypotheses about the relations between UBS Academic Engagement and BIMAS-2 Academic Functioning and Cognitive/Attention subscales, the ratios of false-negative rates to false-positives are concerning given the priority of over-identification at this level of screening. Thus, it is recommended schools do not solely rely on this UBS subscale as a way of identifying students at risk for concerns related to attention, impulsivity, hyperactivity, or academic functioning. Instead, schools are encouraged to use other well-validated screening measures for identifying students with these concerns.

Classification accuracy indices (i.e., sensitivity, specificity, PPV, and NPV) for the UBS Social/Emotional Engagement subscale were partially consistent with hypotheses. First, as speculated, for the BIMAS-2 subscales of Conduct, Negative Affect, and Social, the most favorable accuracy indices were found for UBS Social/Emotional Engagement compared to UBS Academic Readiness and Total scale. Across these three BIMAS-2 subscales, UBS

Social/Emotional Engagement sensitivity values were 38.71% (Conduct), 28.00% (Negative Affect), and 35.00% (Social), suggesting false negative error rates of 61.29%, 72.00%, and 65.00%, respectively. Although specificity, PPV, and NPV values were all within an acceptable range across these analyses, the sensitivity values suggest poor ability of the UBS to predict students presenting with these concerns. As such, low scores on this subscale should not be interpreted as an indication a student is not struggling or does not need support with issues related to social functioning, conduct, or mood and emotions. It is prudent stakeholders continue to use other well-validated screening measures either in concert with or in place of the UBS Social/Emotional Engagement subscale until more research is conducted on the validity of this subscale. Lane and colleagues' (2012) Student Risk Screening Scale – Internalizing and Externalizing [SRSS-IE] version is one example of a well-validated screening measure designed to assess concerns related to both disruptive behavior and mood/anxiety, and may be one option for schools to consider. Interestingly, and contrary to hypotheses, the largest sensitivity value for UBS Social/Emotional Engagement (i.e., 47.06%) was found for BIMAS-2 Cognitive/Attention risk status, which is the BIMAS-2 subscale found to have the largest effect sizes in differentiating samples of children with ADHD (McDougal et al., 2011). Unlike the relation between UBS Academic Readiness and BIMAS-2 Cognitive/Attention, there is not a high degree of overlap between the content of items between UBS Social/Emotional Engagement and the Cognitive/Attention subscale. This may suggest teachers in this sample were influenced by students' ADHD-related behaviors when rating social functioning items. In a study of elementary teacher ratings of student functioning by McConaughy and colleagues (2011), ADHD diagnosis accounted for the most variance in teacher ratings of social skills and school adaptive behaviors. Although diagnostic information is not available for the current sample, the

Cognitive/Attention subscale evidenced the highest frequency of students endorsed as *some* or *high risk* at both administrations (Q2: 24.9%, $n = 51$; Q4: 23.7%, $n = 47$) of all BIMAS-2 subscales. Taken together, there may be some relation between these three constructs (i.e., ADHD, social skills, and academic functioning) in the current sample that could be influencing the strength of relations in the current study.

Overall, both convergent and concurrent criterion validity findings for UBS Social/Emotional Engagement suggest a lack of clear differentiation in the performance of UBS Social/Emotional Engagement with regards to hypothesized (i.e., Conduct, Negative Affect, Social) versus non-hypothesized (i.e., Cognitive/Attention, Academic Functioning) BIMAS-2 subscales. Based on results of these analyses with BIMAS-2 as the criterion, the UBS appears to classify students more accurately for current concerns related to attention, hyperactivity, impulsivity, executive functioning, following directions, and academic performance (i.e., Cognitive/Attention and Academic Functioning subscales). Alternatively, both UBS subscales and the Total scale did not seem to perform as well in accurately classifying students presenting with concerns such as (a) aggression, defiance, risky behaviors, and anger (i.e., Conduct subscale); (b) depression/sadness, anxiety/worry, and other mood problems (i.e., Negative Affect subscale); and (c) difficulty with friendships, social cues, and communication (i.e., Social subscale). While more research is needed to draw formal conclusions about the performance of the UBS, at this time, it is not recommended that UBS scores be used for identifying students in need of supports for concerns related to disruptive behavior, mood and anxiety, and social functioning. The findings of this study suggest more research is needed to determine whether the UBS adequately assesses students with these concerns.

Considering these descriptions of the BIMAS-2 subscales, it appears the UBS may be better able to identify students with inattention, hyperactivity/impulsivity, and academic performance concerns than internalizing concerns. While identifying youth with inattention, hyperactivity/impulsivity, and academic performance concerns is important, these students can also be more easily identified by teacher referral and other objective methods used by schools (e.g., ODRs, suspensions). This is because these concerns often result in observable behaviors and/or require increased teacher attention (e.g., frequent prompts to stay on task or follow directions, prompts to stay seated, longer completion time or incomplete work), which can disrupt the learning environment (Greene et al., 2002; Vile Junod et al., 2006). In contrast, students with internalizing concerns are more likely to be quiet, withdrawn, anxious and consequently unlikely to cause classroom disruption and demand teacher attention (Seeley et al., 2014). As such, they are less likely to be identified and referred for services, suggesting these are precisely the students who might benefit most from using standardized screening measures in initial risk identification. While more research is needed to examine the validity of the UBS, the results of this initial examination of concurrent criterion validity suggest concern about the UBS ability to accurately classify students with internalizing and aggression concerns. Additionally, although results were more favorable for inattention, hyperactivity/impulsivity, and academic performance concerns, the overall poor sensitivity values across all UBS subscales and total scale scores for predicting concurrent BIMAS-2 risk status raise caution about the ability of the measure to accurately identify students who might otherwise fall through the cracks using traditional methods of referral. Thus, schools should exercise caution if they continue to use the UBS within their universal screening protocols until further research is conducted supporting the performance of the UBS as a screener. Schools are encouraged to explore other more well-

established screening measures designed to identify youth with internalizing concerns in order to ensure these students do not miss critical opportunities to receive necessary supports to improve their functioning and reduce their risk of future challenges.

Aim 4: UBS Short-Term Predictive Criterion-Related Validity

Results of predictive validity analyses were more disappointing than those found for concurrent validity. In terms of hypothesized relationships between Q1 UBS and Q4 BIMAS-2 subscales, UBS Academic Readiness correlated significantly more with BIMAS-2 subscales of interest (i.e., Cognitive/Attention and Academic Functioning) than other UBS scales, as predicted, providing additional support for convergent validity of the UBS Academic Readiness subscale. Contrary to hypotheses, UBS Social/Emotional Engagement correlated less with BIMAS-2 subscales of interest (i.e., Conduct, Negative Affect, and Social) than UBS Academic Readiness did, but not to a significant degree. While overall results of all correlation and logistic regression analyses were significant, sensitivity indices across all UBS and BIMAS-2 pairwise combinations were concerning. Specifically, UBS Academic Engagement evidenced the highest sensitivity values (i.e., 48.94% for BIMAS-2 Cognitive/Attention risk status with gender included as a covariate, and 31.43% for BIMAS-2 Academic Functioning risk status), neither of which came close to the 75% threshold deemed acceptable in reviews of the literature (Glover & Albers, 2007). Even in the best-case scenario, false negative error rates (i.e., classifying a student as not at risk for inattention, hyperactivity, and executive functioning concerns when they are, in fact, at risk for those concerns) for predictive validity of the UBS subscales was slightly above chance (i.e., 51.1%). Specificity values were high across all UBS and BIMAS-2 combinations (i.e., ranging from 93.38% to 100.00%) and NPV values were above the 75% acceptable threshold (i.e., ranging from 79.78% to 91.41%). However, the ability of the measure to

accurately identify students at risk of developing social, emotional, and/or behavioral concerns is central to the purpose of this screening measure. Thus, these results suggest the UBS may be a poor predictor of student end-of-year risk status, as indicated by BIMAS-2 subscale *T*-score risk status.

Limitations and Future Directions

The results of the current study should be considered within the context of several significant limitations that might impact the generalizability and interpretation of results. These limitations broadly fall under the domains of: (a) issues with generalizability, (b) concerns related to analytic strategy and the quality of available data, and (c) issues specifically related to the examination of validity in the current study (i.e., criterion measure, types of validity assessed).

Regarding generalizability, this study represents the first investigation of the UBS and thus, findings are preliminary and additional evaluation and replication are recommended. This study involved one small public elementary school in Hawai‘i. While the students sampled in this study were quite similar to the larger HIDOE student population in terms of demographic characteristics (i.e., ethnicity, socioeconomic status, use of special education services, gender), there were fewer students identified as English language learners in the current study compared to the larger HIDOE population (i.e., 3.0% compared to 20.5% statewide; CRDC, n.d.). Additionally, there may be other differences related to teacher characteristics, culture, student academic performance, disciplinary actions, or other factors that were not measured in this study, that could influence the generalizability of these findings. Thus, results can only be interpreted within the context of the school used in this study. Additionally, both my target and criterion measures were teacher-report questionnaires, which can be subject to rating bias and other

confounds that could limit the validity of ratings. Along these lines, future concurrent and predictive validity studies should consider comparison against more objective measures of student functioning, such as ODRs, attendance records, measures of academic functioning, and utilization of counseling services.

With regard to analytic strategy, several limitations are noteworthy. First, while the sample size provided sufficient power for analyses included in this study, it was not large enough to allow for more complicated analyses involving covariates or proper consideration of the nested nature of the data. All students were nested within teacher, which were nested within grade level of a single school. Such data dependencies could have accounted for significant variance within my models but were not accounted for in the current study. One study of a well-established universal screening measure (i.e., BESS; Kamphaus et al., 2007) found that taken together, student- and teacher/classroom-level variables explained more than half of the variance in students' scores on the measure. Future studies with larger sample sizes are needed to examine the influence of similar covariates on UBS scores using more complex analyses (e.g., multilevel modeling) to control for the nested nature of the data and allow for the exploration of these additional variables known to influence student scores on teacher rating scales, such as student and teacher gender (Dowdy et al., 2013; Splett et al., 2018) and race (Blake et al., 2016; Downey & Pribesh, 2004; Saft & Pianta, 2001); student SES (Phillips & Lonigan, 2010); and other teacher-related variables such as years of experience (Mashburn et al., 2006), professional development experiences (Splett et al., 2018), or perceptions of school organizational health (Pas & Bradshaw, 2014). Students in this study also spanned Kindergarten through 6th grade, which encompasses a fairly large range in terms of childhood development. While teachers were asked to rate their students on UBS items relative to same-aged peers, it is possible that some

differences in student ratings could be due to normative differences in developmental abilities between age groups. Thus, if schools wish to continue pursuing use of the UBS, future studies should examine the impact of age on UBS scores and potentially identify different cut off scores (i.e., age-based norms) to use for screening purposes. Additionally, studies may wish to consider accounting for student utilization of Tier 2 and Tier 3 supports and the degree to which students received SEL curriculum throughout the academic year. For example, intersecting with the recommendation above for comparing UBS scores against ODRs, there is a substantial body of evidence showing racial and ethnic disparities in school discipline (Blake et al., 2016; Ladson-Billing, 2006; Losen et al., 2015), and forthcoming studies could examine UBS scores' relations with ODRs as moderated by race or ethnicity (David et al., 2019). Related to SEL curriculum, it is unclear the degree to which students in the current study received *MindUp* (The Hawa Foundation, 2011) throughout the course of the year as this was not something that was systematically monitored by the school. Schools and researchers should consider formally tracking implementation of any intervention delivered to (a) ensure students receive the intervention as directed and (b) be able to link dosage and fidelity of interventions to student outcomes. Regarding the UBS, future studies should investigate the effects of such interventions across universal and other tiers of support, thereby allowing a better understanding of test-retest reliability and the measure's sensitivity to change.

Next, several statistical findings could have been artificially inflated or deflated by aspects of the data. First, data that are non-normal can lead to bias in estimating Cronbach's alpha coefficients (Sheng & Sheng, 2012) and according to Liu and Zumbo (2007), asymmetric outliers can artificially inflate estimates of alpha if their impact is large. While the overall scope of outliers, skewness, and kurtosis was not particularly substantial for the current study, it is

worth noting there is a small possibility these data characteristics could have played a role in the large alpha coefficients found for the UBS. Second, it is important to note this study involved over 100 comparisons run with the same data set. While most analyses were planned, conducting so many significance tests on the same data set leads to alpha inflation (i.e., increasing the chance of false positive error much greater than the acceptable 5%). I attempted to control for this through utilizing Bonferroni corrections in analyses when available in SPSS (i.e., 45 ANOVA post hoc comparisons) and evaluating other analyses against Bonferroni adjusted significance values. However, the potential for false positives in my analyses is still present, particularly for any analyses with significance values closer to $p = .05$ without Bonferroni correction. Third, it is possible that unequal group sizes could have impacted the classification accuracy estimates associated with logistic regression analyses. However, the chance of unequal groups playing a large role in the UBS misclassification of cases seems relatively low as equal groups is not a requirement for logistic regression. Additionally, research noting the impact of unequal groups has focused on much more substantial disproportionality (i.e., one group comprising 1% or less in a sample; King & Zeng, 2001) than was found for the most disproportionate comparison in the current study (i.e., 8.6% [$n = 17$] of sample in the *concern* group for BIMAS-2 Social at Q4). The most likely effect of unequal group size in the current study was reduction of power in the logistic regression models. As such, future studies should consider the potential impact of unequal groups in determining desired sample sizes, particularly if they intend to incorporate multiple predictors in the model.

Lastly, there are several limitations related to my examination of UBS validity, which are perhaps, the most important considerations for nuancing interpretations of this study's results. First, I will note factors which may have influenced the accurate assessment of construct validity.

One consideration is the unknown extent to which the conceptual domain being measured was adequately defined and the content validity of the UBS was adequately established during initial scale development. The UBS was designed to align with CASEL (2015) SEL competencies and HIDEOE GLOs. While theoretical bases for SEL-related items exist in the literature and were used in generating initial UBS items, there is some criticism about a lack of consensus in the field's definition of SEL and relevant behaviors and outcomes (Wigelsworth, 2010). Further, the theoretical foundation of HIDEOE GLOs is not well-established. These issues could potentially invite confusion about what the construct does or does not refer to and thus, how it should relate to other known measures and outcomes. The UBS could benefit from more in-depth examinations of content validity, even potentially revisiting past data with the original larger item pool and using statistical methods (e.g., EFA, Item-Response Theory methods) to better understand the constructs being assessed, and statistically refine the measure. This is important to note here as any limitations in the scale development process could have weakened the psychometric results obtained in this study.

Furthermore, regarding the examination of construct validity, I only examined the UBS in relation to BIMAS-2 risk classification and did not evaluate the UBS in terms of its usability and technical adequacy for monitoring students' progress across SEL and GLO competencies. Given this is the other intended purpose of the UBS, additional studies are needed to evaluate the extent to which the UBS adequately measures SEL and HIDEOE GLO constructs. The factor analysis provided some support for differentiation between the two factors with the first factor seeming to align more with SEL competencies and the second factor appearing to incorporate GLO competencies. However, future studies may wish to examine the UBS for these purposes, outside of its use as a universal screening tool.

Lastly, one of the most significant limitations of this study is in the use of the BIMAS-2 as the reference standard for examinations of convergent and criterion-related validity. Overall, the psychometric performance of the BIMAS-2 as a universal screening measure of social, emotional, and behavioral health is not well-established. Measure developers examined and reported various psychometric properties of the BIMAS-2 in their technical manual; however, most analyses appear to be conducted with the same sample and results have not been subject to independent replication or peer review. Additionally, no local norms (e.g., school-, district-, state-level) exist for the BIMAS-2 so it is unknown how this measure performs with the population from which the current study sampled. The BIMAS-2 was used in the current study as the criterion measure for practical purposes, as this was the measure ordained for use by the school for universal screening and progress monitoring. However, it is important to acknowledge the BIMAS-2 may not represent an adequate “reference standard” of the target condition, as is typically used in examinations of criterion validity (Trevethan, 2017). As such, we cannot extrapolate the results past this comparison with the BIMAS-2. More research is needed with other well-established criterion measures to draw inferences about how well the UBS actually does in classifying students who are known to have social, emotional, and/or behavioral concerns or predicting those who go on to develop them. Several well-validated measures were identified throughout the process of examining the UBS, including a few free, relatively brief screeners (e.g., SRSS-IE, Lane, Oakes, et al., 2012; Lane et al., 2015; Strengths and Difficulties Questionnaire [SDQ], Goodman, 1997). Given the aforementioned limitations of the BIMAS-2, future studies may consider comparing the UBS to performance of either one of these well-established screening measures to allow for a better comparison.

Study Implications

This is the first study to examine the psychometric properties of the UBS, a novel, locally developed SEL-focused universal screening tool designed to identify students at risk for social, emotional, and/or behavioral concerns. This measure aligns well with local needs and has high usability, particularly related to the feasibility of its use given the strong buy in from school staff (e.g., 100% completion rate), no financial cost, and low time burden on staff. Taking into account the aforementioned limitations, this initial examination found mixed results about the technical adequacy of the measure with the current sample, including: some evidence supporting factor validity of the UBS through the identified factor structure, adequate reliability for UBS total scale and subscale scores, and general convergence between UBS and BIMAS-2 subscale scores, but notably poor evidence for UBS sensitivity in classifying and predicting students presenting with social, emotional, behavioral, and/or academic concerns when determined by BIMAS-2 risk status.

From a practical standpoint, this study aimed to assist the HDOE in answering the question: *Can the UBS be used in place of the BIMAS-2 as an initial assessment of student needs in a multi-gate screening approach?* While more research is needed to conclusively answer this question, the initial results were not promising. Overall, students' scores on the two measures were strongly related to each other and there was some evidence suggesting potential redundancy between the UBS Academic Engagement subscale and both the BIMAS-2 Cognitive/Attention and Academic Functioning subscales when administered concurrently. However, classification accuracy indices for both concurrent and predictive validity did not seem to suggest adequate overlap in risk classification between the measures. Schools wanting to use the UBS to identify students are encouraged to use other well-validated screening measures and objective data in


combination with the UBS for risk identification until more research is conducted evaluating the ability of the UBS to accurately classify students at risk of social, emotional, and/or behavioral concerns.





Appendix A

BIMAS-2, Teacher Standard Form



Teacher Standard Form (Ages 5 – 18)

Shade circles like this: 

Not like this:    

James L. McDougal, Psy.D., Achilles N. Bardos, Ph.D., & Scott T. Meier, Ph.D.

Rating:

During the past week, this student...

- ① = Never (0 times or not observed)
- ② = Rarely (Observed 1-2 times or to a minimal extent)
- ③ = Sometimes (Observed 3-4 times or to a moderate extent)
- ④ = Often (Observed 5-6 times or to a significant extent)
- ⑤ = Very Often (Observed 7 or more times or to an extreme extent)

<i>During the past week, this student...</i>	Never	Rarely	Sometimes	Often	Very Often
1. shared what they were thinking about.	①	②	③	④	⑤
2. appeared angry.	①	②	③	④	⑤
3. had trouble paying attention.	①	②	③	④	⑤
4. followed directions.	①	②	③	④	⑤
5. appeared sleepy or tired.	①	②	③	④	⑤
6. was impulsive.	①	②	③	④	⑤
7. spoke clearly with others.	①	②	③	④	⑤
8. appeared depressed.	①	②	③	④	⑤
9. engaged in risk-taking behavior.	①	②	③	④	⑤
10. had problems staying on task.	①	②	③	④	⑤
11. maintained friendships.	①	②	③	④	⑤
12. acted sad or withdrawn.	①	②	③	④	⑤
13. fought with others (verbally, physically, or both).	①	②	③	④	⑤
14. acted without thinking.	①	②	③	④	⑤
15. appeared comfortable when relating to others.	①	②	③	④	⑤
16. was easily embarrassed or felt ashamed.	①	②	③	④	⑤
17. lied or cheated.	①	②	③	④	⑤
18. had trouble remembering.	①	②	③	④	⑤
19. was generally friendly with others.	①	②	③	④	⑤
20. appeared anxious (worried or nervous).	①	②	③	④	⑤
21. lost his/her temper when upset.	①	②	③	④	⑤
22. had trouble with organizing and planning.	①	②	③	④	⑤
23. worked out problems with others.	①	②	③	④	⑤
24. expressed thoughts of hurting himself/herself.	①	②	③	④	⑤
25. was aggressive (threatened or bullied others).	①	②	③	④	⑤
26. received failing grades at school.	①	②	③	④	⑤
27. was emotional or upset.	①	②	③	④	⑤
28. fidgeted.	①	②	③	④	⑤
29. was suspected of using alcohol and/or drugs.	①	②	③	④	⑤
30. worked up to his/her academic potential.	①	②	③	④	⑤
31. was sent to an authority for discipline.	①	②	③	④	⑤
32. was suspected of smoking or chewing tobacco.	①	②	③	④	⑤
33. was prepared for class.	①	②	③	④	⑤
34. was absent from school.	①	②	③	④	⑤

**Thank you for completing this questionnaire.
Please make sure you have answered every item.**

Copyright © 2016 Edumetrix, LLC. All rights reserved. In the U.S.A., P.O. Box 336404 Greeley, CO 80633 (970) 301-5166

Appendix B

BIMAS-2 Subscale Legend

BEHAVIORAL CONCERN SCALES

CONDUCT	
2	appeared angry
9	engaged in risk-taking behavior
13	fought with others (verbally, physically, or both)
17	lied or cheated
21	lost their temper when upset
25	was aggressive (threatened or bullied others)
29	was suspected of using alcohol and/or drugs
31	was sent to an authority for discipline
32	was suspected of smoking or chewing tobacco

NEGATIVE AFFECT

5	appeared sleepy or tired
8	appeared depressed
12	acted sad or withdrawn
16	was easily embarrassed or felt ashamed
20	appeared anxious (worried or nervous)
24	expressed thoughts of hurting themselves
27	was emotional or upset

COGNITIVE/ATTENTION

3	had trouble paying attention
6	was impulsive
10	had problems staying on task
14	acted without thinking
18	had trouble remembering
22	had trouble with organizing and planning
28	fidgeted

ADAPTIVE SCALES

SOCIAL	
1	shared what they were thinking about
7	spoke clearly with others
11	maintained friendships
15	appeared comfortable when relating to others
19	was generally friendly with others
23	worked out problems with others

ACADEMIC FUNCTIONING

4	followed directions
26	received failing grades at school
30	worked up to their academic potential
33	was prepared for class
34	was absent from school

Appendix C

Item-Level Statistics Table for BIMAS-2 Q2 (*n* = 205) and Q4 (*n* = 211)

Item	Q2 Statistics			Q4 BIMAS-2 Statistics		
	<i>M</i>	<i>SD</i>	<i>r</i> with total	<i>M</i>	<i>SD</i>	<i>r</i> with total
Conduct Items						
2. Appeared angry	0.51	0.80	.64	0.46	0.83	.66
9. Engaged in risk-taking behavior	0.37	0.75	.37	0.44	0.88	.20
13. Fought with others	0.47	0.78	.71	0.40	0.78	.73
17. Lied or cheated	0.55	0.88	.50	0.39	0.78	.62
21. Lost their temper when upset	0.31	0.72	.75	0.28	0.73	.74
25. Was aggressive	0.11	0.42	.66	0.15	0.53	.67
29. Was suspected of using alcohol or drugs ^a	0.00	0.00		0.00	0.00	
31. Was sent to an authority for discipline	0.12	0.47	.55	0.14	0.47	.60
32. Was suspected of smoking/chewing tobacco ^b	0.00	0.00		0.01	0.10	.13
Negative Affect Items						
5. Appeared sleepy or tired	0.88	0.99	.49	0.78	1.02	.51
8. Appeared depressed	0.37	0.68	.75	0.27	0.61	.67
12. Acted sad or withdrawn	0.56	0.87	.72	0.37	0.68	.69
16. Was easily embarrassed or felt ashamed	0.92	1.04	.59	0.59	0.79	.35
20. Appeared anxious	0.65	0.89	.72	0.43	0.74	.47
24. Expressed thoughts of hurting themselves	0.14	0.59	.19	0.04	0.30	.28
27. Was emotional or upset	0.66	0.95	.70	0.49	0.86	.59
Cognitive/Attention Items						
3. Had trouble paying attention	1.86	1.23	.81	1.62	1.30	.85
6. Was impulsive	0.92	1.26	.77	0.82	1.25	.71
10. Had problems staying on task	1.66	1.29	.85	1.40	1.34	.82
14. Acted without thinking	1.14	1.24	.80	0.92	1.20	.76

Item	Q2 BIMAS-2 Statistics			Q4 BIMAS-2 Statistics		
	<i>M</i>	<i>SD</i>	<i>r</i> with total	<i>M</i>	<i>SD</i>	<i>r</i> with total
Cognitive/Attention Items (Continued)						
18. Had trouble remembering	1.02	1.18	.69	0.60	0.98	.58
22. Had trouble with organizing and planning	1.42	1.36	.75	1.04	1.30	.69
28. Fidgeted	1.17	1.27	.80	1.06	1.23	.69
Social Items						
1. Shared what they were thinking about	2.73	1.01	.34	2.92	1.04	.25
7. Spoke clearly with others	3.16	0.76	.66	3.36	0.71	.47
11. Maintained friendships	3.21	0.82	.68	3.39	0.78	.56
15. Appeared comfortable when relating to others	3.13	0.83	.64	3.43	0.65	.61
19. Was generally friendly with others	3.30	0.64	.64	3.44	0.70	.47
23. Worked out problems with others	2.44	1.24	.48	2.72	1.34	.24
Academic Functioning						
4. Followed directions	2.76	0.95	.65	2.97	0.95	.63
26. Received failing grades at school	3.31	1.00	.61	3.49	0.90	.68
30. Worked up to their academic potential	2.84	0.89	.73	3.01	0.95	.67
33. Was prepared for class	2.97	0.86	.66	3.21	0.86	.63
34. Was absent from school	3.57	0.82	.25	3.49	0.80	.29

Note. Bolded mean values indicate means that may have a pronounced floor or ceiling effect.

Bolded item-total correlation values indicate coefficients meeting the criterion of $r < .35$ (Lane et al., 2016; Walker et al., 2009), suggesting potentially questionable consistency with the construct being measured by the subscale.

^a Item 29 had zero variance for both Q2 and Q4 administrations (i.e., no students were rated > 0).

^b Item 32 had zero variance for the Q2 administration (i.e., no students were rated > 0).

Appendix D: Universal Behavior Screener

Universal Behavior Screener

Teacher: _____

Session: QTR 1 QTR 2 QTR 3 QTR 4

Listed below are some behaviors that describe how students may act. Please review each behavior and then rate each student (including students with an IEP) in your class on a scale of 1-5 based off of your observations in the past MONTH on what you expect of a typical child in your class. Don't worry about comparing your ratings to your peers – data will be based on student growth based on own teacher's rating. Add new students to the bottom of the list. If a student has left, leave the ratings blank. Put a "star" next to the top three students you are concerned about their behavior at school.

- # 1: Cooperates with Peers – Plays and works well with others, is kind, is a community contributor
- # 2: Uses Socially Appropriate Responses – Responds appropriately to the emotions/behaviors of others, is an effective communicator
- # 3: Is Prepared to Learn – Arrives on time, is responsible for school materials, is a self-directed learner
- # 4: Engages in Academic Tasks – Starts, works, & finishes academic tasks within reasonable time frames, is a quality producer
- # 5: Follows Rules, Routines, Directions – Responds safely to expectations and/or changes in the environment
- # 6: Has a Positive Attitude – Demonstrates a "growth" mindset, is interested in improving the school community
- # 7: Regulates Emotions – Does not demonstrate intense feelings of sadness, worry, anger, etc.
- # 8: Is an Effective Problem Solver – Puts thought into decisions (i.e., not impulsive), is a complex thinker, ethical user of technology
- # 9: Pays Attention – Is focused & is not overly distracted

Write 1 if the student is **Deficient** in performing the expectation or if the behavior never occurs

Write 2 if the student is **Well Below** performing the expectation or if the behavior rarely occurs

Write 3 if the student is **Developing Proficiency** by approaching acceptable achievement of the expectation

Write 4 if the student **Meets Proficiency** in demonstrating acceptable achievement of the expectation

Write 5 if the student **Meets with Excellence** in performing the expectation above the typical/average student

Student Name	#1	#2	#3	#4	#5	#6	#7	#8	#9
1.									
2.									
3.									
4.									
5.									
6.									
7.									
8.									
9.									
10.									

Appendix E: Item-Level Statistics Tables for UBS Q1 – Q4

Table E1

Q1 UBS Item-Level Statistics, Corrected Item-Total Correlations, and Cronbach's Coefficient Alphas if Item Deleted Statistics (n = 214)

UBS Item	<i>M</i>	<i>SD</i>	Skew	Kurtosis	Item-total <i>r</i>	<i>α</i>
Factor 1: Social/Emotional Engagement						.94
1. Cooperates with peers	3.48	0.81	-0.70	0.80	.82	.93
2. Uses socially appropriate responses	3.42	0.84	-0.57	0.81	.86	.92
5. Follows rules, routines, and directions	3.55	0.73	-1.20	2.02	.80	.93
6. Has a positive attitude	3.59	0.77	-0.91	1.32	.84	.93
7. Regulates emotions	3.51	0.83	-0.81	0.72	.87	.92
Factor 2: Academic Readiness						.91
3. Is prepared to learn	3.40	0.83	-0.72	0.59	.82	.88
4. Engages in academic tasks	3.26	0.89	-0.36	-0.39	.79	.89
8. Is an effective problem solver	3.23	0.82	-0.65	-0.13	.81	.88
9. Pays attention	3.08	0.96	-0.30	-0.50	.79	.89
Total (All 9 items)						.95
1. Cooperates with peers					.79	.95
2. Uses socially appropriate responses					.84	.95
3. Is prepared to learn					.86	.94
4. Engages in academic tasks					.75	.95
5. Follows rules, routines, & directions					.84	.95
6. Has a positive attitude					.82	.95
7. Regulates emotions					.83	.95
8. Is an effective problem solver					.79	.95
9. Pays attention					.78	.95

Note. All data from Q1. The Cronbach's alpha coefficients reported at item-level are *alpha if item removed* values, whereas alpha coefficients reported for each factor are the factor-level Cronbach's alpha values.

Table E2

Q3 UBS Item-Level Statistics, Corrected Item-Total Correlations, and Cronbach's Coefficient Alphas if Item Deleted Statistics (n = 217)

UBS Item	<i>M</i>	<i>SD</i>	Skew	Kurtosis	Item-total <i>r</i>	α
Factor 1: Social/Emotional Engagement						.92
1. Cooperates with peers	3.74	0.76	-0.73	0.77	.84	.90
2. Uses socially appropriate responses	3.62	0.78	-0.78	0.39	.80	.90
5. Follows rules, routines, & directions	3.72	0.77	-0.52	0.43	.74	.92
6. Has a positive attitude	3.84	0.64	-0.98	2.46	.81	.90
7. Regulates emotions	3.81	0.76	-0.96	1.31	.80	.90
Factor 2: Academic Readiness						.92
3. Is prepared to learn	3.62	0.82	-0.55	0.51	.82	.89
4. Engages in academic tasks	3.69	0.79	-0.62	0.65	.80	.89
8. Is an effective problem solver	3.50	0.79	-0.48	0.15	.78	.90
9. Pays attention	3.48	0.82	-0.39	0.20	.83	.88
Total (All 9 items)						.95
1. Cooperates with peers					.80	.94
2. Uses socially appropriate responses					.80	.94
3. Is prepared to learn					.80	.94
4. Engages in academic tasks					.80	.94
5. Follows rules, routines, & directions					.82	.94
6. Has a positive attitude					.76	.94
7. Regulates emotions					.75	.94
8. Is an effective problem solver					.79	.94
9. Pays attention					.79	.94

Note. All data from Q3. The Cronbach's alpha coefficients reported at item-level are *alpha if item removed* values, whereas alpha coefficients reported for each factor are the factor-level Cronbach's alpha values.

Table E3

Q4 UBS Item-Level Statistics, Corrected Item-Total Correlations, and Cronbach's Coefficient Alphas if Item Deleted Statistics (n = 213)

UBS Item	<i>M</i>	<i>SD</i>	Skew	Kurtosis	Item-total <i>r</i>	<i>α</i>
Factor 1: Social/Emotional Engagement						.94
1. Cooperates with peers	3.87	0.82	-0.55	0.01	.84	.92
2. Uses socially appropriate responses	3.83	0.80	-0.62	0.48	.83	.92
5. Follows rules, routines, and directions	3.86	0.83	-0.65	0.65	.81	.92
6. Has a positive attitude	4.04	0.74	-0.49	0.12	.82	.92
7. Regulates emotions	3.92	0.90	-1.00	1.31	.85	.92
Factor 2: Academic Readiness						.91
3. Is prepared to learn	3.79	0.85	-0.28	-0.30	.79	.89
4. Engages in academic tasks	3.83	0.85	-0.55	0.52	.84	.87
8. Is an effective problem solver	3.68	0.82	-0.81	1.34	.73	.90
9. Pays attention	3.67	0.89	-0.49	0.12	.82	.87
Total (All 9 items)						.95
1. Cooperates with peers					.81	.95
2. Uses socially appropriate responses					.82	.94
3. Is prepared to learn					.80	.95
4. Engages in academic tasks					.81	.95
5. Follows rules, routines, & directions					.85	.94
6. Has a positive attitude					.79	.95
7. Regulates emotions					.82	.94
8. Is an effective problem solver					.70	.95
9. Pays attention					.84	.94

Note. All data from Q4. The Cronbach's alpha coefficients reported at item-level are *alpha if item removed* values, whereas alpha coefficients reported for each factor are the factor-level Cronbach's alpha values.

Appendix F: Exploration of Distribution and Scope of UBS & BIMAS-2 Outliers

Using z -Scores

Additional examination of the data focused on assessing the scope and distribution of outliers across UBS and BIMAS-2 subscales and/or total scale to decide as to whether the proposed parametric tests in Aims 2 – 4 remained appropriate. Z -scores were calculated for subscale and/or total scale scores of each measure to statistically discriminate between extreme, probable, and potential outliers and assess the extent to which outliers diverged from expected ranges of a normal distribution. Thresholds of close to 0%, 1%, and 5% were used as expected range of extreme ($|z| > 3.29$), probable ($|z| > 2.58$), and potential ($|z| > 1.96$) outliers, respectively, for normal distribution.

Overall, both UBS subscales at Q2 as well as UBS Factor 2 and Total at Q3 failed to meet the expected 95% of scores falling within the normal range (i.e., $|z| < 1.96$). The percentage of scores falling within the normal range for these scales ranged from 90.6% to 91.5%. Extreme outliers were found across total scale and/or subscales at all time points except Q2, ranging from one to three cases. Q1 Factor 1 subscale and total scale had the greatest number of extreme outliers with three cases each and no extreme outliers were found for total scale or either subscale at Q2. Probable outliers exceeding 1% of scores were found for many of the total scale/subscales across time points, ranging from one to four cases more than what would be expected in normal distribution. Potential outliers exceeding the expected 5% of scores were found at Q2 and Q3, ranging from 14 to 17 cases across total scale and subscales.

Table F1*UBS Scale Outliers Based on z-Scores*

UBS Scale	$z > 3.29$ Extreme outliers	$z > 2.58$ Probable outliers	$z > 1.96$ Potential Outliers	$z < 1.96$ Normal range
Quarter 1 (Q1; $n = 214$)				
UBS Factor 1	1.4% ($n = 3$)	0.9% ($n = 2$)	2.8% ($n = 6$)	94.9% ($n = 203$)
UBS Factor 2	0.0% ($n = 0$)	1.4% ($n = 3$)	3.3% ($n = 7$)	95.3% ($n = 204$)
UBS Total	1.4% ($n = 3$)	0.9% ($n = 2$)	2.8% ($n = 6$)	94.9% ($n = 203$)
Quarter 2 (Q2; $n = 213$)				
UBS Factor 1	0.0% ($n = 0$)	2.8% ($n = 6$)	6.6% ($n = 14$)	90.6% ($n = 193$)
UBS Factor 2	0.0% ($n = 0$)	0.9% ($n = 2$)	7.5% ($n = 16$)	91.5% ($n = 195$)
UBS Total	0.0% ($n = 0$)	0.9% ($n = 2$)	4.2% ($n = 9$)	94.8% ($n = 202$)
Quarter 3 (Q3; $n = 218$)				
UBS Factor 1	0.5% ($n = 1$)	1.8% ($n = 4$)	2.3% ($n = 5$)	95.4% ($n = 207$)
UBS Factor 2	0.5% ($n = 1$)	0.5% ($n = 1$)	7.8% ($n = 17$)	91.2% ($n = 198$)
UBS Total	0.5% ($n = 1$)	1.4% ($n = 3$)	6.9% ($n = 15$)	91.2% ($n = 198$)
Quarter 4 (Q4; $n = 214$)				
UBS Factor 1	0.5% ($n = 1$)	1.4% ($n = 3$)	2.8% ($n = 6$)	95.3% ($n = 203$)
UBS Factor 2	0.5% ($n = 1$)	1.4% ($n = 3$)	2.3% ($n = 5$)	95.8% ($n = 204$)
UBS Total	0.9% ($n = 2$)	1.4% ($n = 3$)	1.4% ($n = 3$)	96.2% ($n = 205$)

Note. Bolded cells indicate percentages outside of expected range for normal distribution.

Expected percentage for extreme, probable, and potential outliers are ~0%, ~1%, and ~5% of scores, respectively. Expected percentage of sample falling in normal range is ~95%.

Closer examination was conducted using z -scores to statistically discriminate between extreme ($|z| > 3.29$), probable ($|z| > 2.58$), and potential ($|z| > 1.96$) outliers and assess the extent to which outliers diverged from expected ranges of a normal distribution (see Table F2 below). All subscales across both time points met the expected 95% of scores in the normal range (i.e., $|z|$

< 1.96). Closer inspection showed only one subscale (i.e., Q4 Conduct) exceeded an expected range of extreme outliers. Three subscales (Q2 Conduct and Academic functioning, Q4 Academic Functioning) exceeded the expected range for probable outliers spanning from one to three cases greater than expected. Only one subscale (i.e., Q2 Social) exceeded the expected 5% of cases showing potential outliers, with 12 cases more than what would be expected for a normal distribution.

Table F2

BIMAS-2 Scale Outliers Based on z-Scores

UBS Scale	$z > 3.29$ Extreme outliers	$z > 2.58$ Probable outliers	$z > 1.96$ Potential Outliers	$z < 1.96$ Normal range
Quarter 2 (Q2; $n = 205$)				
Conduct	0.5% ($n = 1$)	2.4% ($n = 5$)	1.5% ($n = 3$)	95.6% ($n = 196$)
Negative Affect	0.0% ($n = 0$)	1.0% ($n = 2$)	3.4% ($n = 7$)	95.6% ($n = 196$)
Cognitive/Attention	0.0% ($n = 0$)	0.0% ($n = 0$)	2.9% ($n = 6$)	97.1% ($n = 199$)
Social	0.0% ($n = 0$)	0.0% ($n = 0$)	8.3% ($n = 17$)	91.7% ($n = 188$)
Academic Functioning	0.0% ($n = 0$)	2.0% ($n = 4$)	1.0% ($n = 2$)	97.1% ($n = 199$)
Quarter 4 (Q4; $n = 211$)				
Conduct	1.4% ($n = 3$)	0.5% ($n = 1$)	1.4% ($n = 3$)	96.7% ($n = 204$)
Negative Affect	0.5% ($n = 1$)	0.9% ($n = 2$)	1.4% ($n = 3$)	97.2% ($n = 205$)
Cognitive/Attention	0.0% ($n = 0$)	0.0% ($n = 0$)	4.7% ($n = 10$)	95.3% ($n = 201$)
Social	0.0% ($n = 0$)	0.0% ($n = 0$)	2.4% ($n = 5$)	97.6% ($n = 206$)
Academic Functioning	0.0% ($n = 0$)	1.4% ($n = 3$)	1.4% ($n = 3$)	97.2% ($n = 205$)

Note. Bolded cells indicate percentages outside of expected range for normal distribution.

Expected percentage for extreme, probable, and potential outliers are ~0%, ~1%, and ~5%, respectively. Expected percentage of sample falling in normal range is ~95%.

Appendix G

Aim 3 Logistic Regression and Classification Tables:

Q2 UBS Social/Emotional Engagement Subscale and Q2 BIMAS-2 Subscales

Table G1

Logistic Regression Analysis Q2 UBS Social/Emotional Engagement predicting Q2 BIMAS-2 Conduct Subscale Risk Status Membership (n = 205)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^{β} (odds ratio)
UBS Social/Emotional	-2.339	.389	36.114	1	< .001	.096
Constant	6.311	1.291	23.916	1	< .001	550.823
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			53.031	1	< .001	
Score test			56.046	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			10.180	6	.117	

Note. Cox & Snell $R^2 = .228$. Nagelkerke $R^2 = .398$. All statistics herein use three decimal places to maintain precision.

Table G2

Q2 UBS Social/Emotional Engagement Subscale: Classification Table for Q2 BIMAS-2 Conduct Risk Status

Observed	Predicted		% Correct
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	12	19	38.71%
No (Low Risk)	6	168	96.55%
Overall % Correct			87.80%
Change in % Correct from Null Model			2.9%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table G3

Logistic Regression Analysis Q2 UBS Social/Emotional Engagement predicting Q2 BIMAS-2 Negative Affect Subscale Risk Status Membership (n = 205)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Social/Emotional	-1.634	.298	30.066	1	< .001	.195
Constant	4.726	1.061	19.842	1	< .001	112.805
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			37.859	1	< .001	
Score test			38.341	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			2.961	6	.814	

Note. Cox & Snell $R^2 = .169$. Nagelkerke $R^2 = .251$. All statistics herein use three decimal places to maintain precision.

Table G4

Q2 UBS Social/Emotional Engagement Subscale: Classification Table for Q2 BIMAS-2 Negative Affect Risk Status

Observed	Predicted		% Correct
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	14	36	28.00%
No (Low Risk)	8	147	94.84%
Overall % Correct			78.54%
Change in % Correct from Null Model			2.9%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table G5

Logistic Regression Analysis Q2 UBS Social/Emotional Engagement predicting Q2 BIMAS-2 Cognitive/Attention Subscale Risk Status Membership (n = 205)

Predictor	β	$SE \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Social/Emotional	-2.533	.388	42.678	1	< .001	.079
Constant	7.909	1.370	33.333	1	< .001	2720.788
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			70.446	1	< .001	
Score test			66.971	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			13.747	6	.033	

Note. Cox & Snell $R^2 = .291$. Nagelkerke $R^2 = .431$. All statistics herein use three decimal places to maintain precision.

Table G6

Q2 UBS Social/Emotional Engagement Subscale: Classification Table for Q2 BIMAS-2 Cognitive/Attention Risk Status

Observed	Predicted		% Correct
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	24	27	47.06%
No (Low Risk)	9	145	94.16%
Overall % Correct			82.44%
Change in % Correct from Null Model			7.3%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table G7

Logistic Regression Analysis Q2 UBS Social/Emotional Engagement predicting Q2 BIMAS-2 Social Subscale Risk Status Membership (n = 205)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Social/Emotional	-1.647	.307	28.789	1	< .001	.193
Constant	4.415	1.069	17.071	1	< .001	82.697
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			35.139	1	< .001	
Score test			36.813	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			12.185	6	.058	

Note. Cox & Snell $R^2 = .158$. Nagelkerke $R^2 = .251$. All statistics herein use three decimal places to maintain precision.

Table G8

Q2 UBS Social/Emotional Engagement Subscale: Classification Table for Q2 BIMAS-2 Social Risk Status

Observed	Predicted		% Correct
	Yes (Concern)	No (No Concern)	
Yes (Concern)	14	26	35.00%
No (No Concern)	4	161	97.58%
Overall % Correct			85.37%
Change in % Correct from Null Model			4.9%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table G9

Logistic Regression Analysis Q2 UBS Social/Emotional Engagement predicting Q2 BIMAS-2 Academic Functioning Subscale Risk Status Membership (n = 205)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Social/Emotional	-2.297	.369	38.828	1	< .001	.101
Constant	6.671	1.269	27.615	1	< .001	788.854
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			58.269	1	< .001	
Score test			58.776	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			10.208	6	.116	

Note. Cox & Snell $R^2 = .247$. Nagelkerke $R^2 = .391$. All statistics herein use three decimal places to maintain precision.

Table G10

Q2 UBS Social/Emotional Engagement Subscale: Classification Table for Q2 BIMAS-2 Academic Functioning Risk Status

Observed	Predicted		% Correct
	Yes (Concern)	No (No Concern)	
Yes (Concern)	14	27	34.15%
No (No Concern)	8	156	95.12%
Overall % Correct			82.93%
Change in % Correct from Null Model			2.9%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Appendix H

Aim 3 Logistic Regression and Classification Tables: Q2 UBS Academic Readiness Subscale and Q2 BIMAS-2 Subscales

Table H1

Logistic Regression Analysis Q2 UBS Academic Readiness predicting Q2 BIMAS-2 Conduct Subscale Risk Status Membership (n = 205)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^{β} (odds ratio)
UBS Academic Readiness	-1.615	.313	26.610	1	< .001	.199
Constant	3.489	.968	12.994	1	< .001	32.760
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			33.834	1	< .001	
Score test			33.777	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			8.572	7	.285	

Note. Cox & Snell $R^2 = .152$. Nagelkerke $R^2 = .266$. All statistics herein use three decimal places to maintain precision.

Table H2

Q2 UBS Academic Readiness Subscale: Classification Table for Q2 BIMAS-2 Conduct Risk Status

Observed	Predicted		% Correct
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	8	23	25.81%
No (Low Risk)	4	170	97.70%
Overall % Correct			86.83%
Change in % Correct from Null Model			1.9%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table H3

Logistic Regression Analysis Q2 UBS Academic Readiness predicting Q2 BIMAS-2 Negative Affect Subscale Risk Status Membership (n = 205)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Academic Readiness	-1.252	.252	24.601	1	< .001	.286
Constant	3.072	.837	13.474	1	< .001	21.576
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			29.433	1	< .001	
Score test			28.923	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			3.813	7	.801	

Note. Cox & Snell $R^2 = .134$. Nagelkerke $R^2 = .199$. All statistics herein use three decimal places to maintain precision.

Table H4

Q2 UBS Academic Readiness Subscale: Classification Table for Q2 BIMAS-2 Negative Affect Risk Status

Observed	Predicted		% Correct
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	11	39	22.00%
No (Low Risk)	8	147	94.84%
Overall % Correct			77.07%
Change in % Correct from Null Model			1.5%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table H5

Logistic Regression Analysis Q2 UBS Academic Readiness predicting Q2 BIMAS-2 Cognitive/Attention Subscale Risk Status Membership (n = 205)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Academic Readiness	-2.713	.409	44.037	1	< .001	.066
Constant	7.761	1.304	35.415	1	< .001	2346.374
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			85.663	1	< .001	
Score test			74.611	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			16.952	7	.018	

Note. Cox & Snell $R^2 = .342$. Nagelkerke $R^2 = .506$. All statistics herein use three decimal places to maintain precision.

Table H6

Q2 UBS Academic Readiness Subscale: Classification Table for Q2 BIMAS-2 Cognitive/Attention Risk Status

Observed	Predicted		% Correct
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	31	20	60.78%
No (Low Risk)	8	146	94.81%
Overall % Correct			86.34%
Change in % Correct from Null Model			11.2%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table H7

Logistic Regression Analysis Q2 UBS Academic Readiness predicting Q2 BIMAS-2 Social Subscale Risk Status Membership (n = 205)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Academic Readiness	-1.165	.260	20.115	1	< .001	.312
Constant	2.460	.847	8.438	1	.004	11.702
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			22.979	1	< .001	
Score test			23.074	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			5.786	7	.565	

Note. Cox & Snell $R^2 = .106$. Nagelkerke $R^2 = .169$. All statistics herein use three decimal places to maintain precision.

Table H8

Q2 UBS Academic Readiness Subscale: Classification Table for Q2 BIMAS-2 Social Risk Status

Observed	Predicted		% Correct
	Yes (Concern)	No (No Concern)	
Yes (Concern)	6	34	15.00%
No (No Concern)	6	159	96.36%
Overall % Correct			80.49%
Change in % Correct from Null Model			0.0%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table H9

Logistic Regression Analysis Q2 UBS Academic Readiness predicting Q2 BIMAS-2 Academic Functioning Subscale Risk Status Membership (n = 205)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Academic Readiness	-4.171	.693	36.176	1	< .001	.015
Constant	11.663	2.085	31.301	1	< .001	116160.743
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			111.827	1	< .001	
Score test			91.127	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			4.704	7	.696	

Note. Cox & Snell $R^2 = .420$. Nagelkerke $R^2 = .665$. All statistics herein use three decimal places to maintain precision.

Table H10

Q2 UBS Academic Readiness Subscale: Classification Table for Q2 BIMAS-2 Academic Functioning Risk Status

Observed	Predicted		% Correct
	Yes (Concern)	No (No Concern)	
Yes (Concern)	30	11	73.17%
No (No Concern)	9	155	94.51%
Overall % Correct			90.24%
Change in % Correct from Null Model			10.2%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Appendix I

Aim 3 Logistic Regression and Classification Tables:

Q2 UBS Total and Q2 BIMAS-2 Subscales

Table I1

Logistic Regression Analysis Q2 UBS Total predicting Q2 BIMAS-2 Conduct Subscale Risk Status Membership (n = 205)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Total	-2.326	.400	33.859	1	< .001	.098
Constant	6.039	1.280	22.265	1	< .001	419.632
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			49.736	1	< .001	
Score test			50.060	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			7.108	6	.311	

Note. Cox & Snell $R^2 = .215$. Nagelkerke $R^2 = .376$. All statistics herein use three decimal places to maintain precision.

Table I2

Q2 UBS Total: Classification Table for Q2 BIMAS-2 Conduct Risk Status

Observed	Predicted		% Correct
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	11	20	35.48%
No (Low Risk)	7	167	95.98%
Overall % Correct			86.83%
Change in % Correct from Null Model			1.9%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table I3

Logistic Regression Analysis Q2 UBS Total predicting Q2 BIMAS-2 Negative Affect Subscale Risk Status Membership (n = 205)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Total	-1.665	.304	30.020	1	< .001	.189
Constant	4.663	1.045	19.918	1	< .001	106.006
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			38.287	1	< .001	
Score test			37.738	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			3.975	6	.680	

Note. Cox & Snell $R^2 = .170$. Nagelkerke $R^2 = .254$. All statistics herein use three decimal places to maintain precision.

Table I4

Q2 UBS Total: Classification Table for Q2 BIMAS-2 Negative Affect Risk Status

Observed	Predicted		%
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	16	34	32.00%
No (Low Risk)	10	145	93.55%
Overall % Correct			78.54%
Change in % Correct from Null Model			2.9%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table I5

Logistic Regression Analysis Q2 UBS Total predicting Q2 BIMAS-2 Cognitive/Attention Subscale Risk Status Membership (n = 205)

Predictor	β	$SE \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Total	-3.197	.474	45.466	1	< .001	.041
Constant	9.848	1.598	37.982	1	< .001	18919.595
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			89.839	1	< .001	
Score test			79.190	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			7.764	6	.256	

Note. Cox & Snell $R^2 = .355$. Nagelkerke $R^2 = .526$. All statistics herein use three decimal places to maintain precision.

Table I6

Q2 UBS Total: Classification Table for Q2 BIMAS-2 Cognitive/Attention Risk Status

Observed	Predicted		% Correct
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	27	24	52.94%
No (Low Risk)	11	143	92.86%
Overall % Correct			82.93%
Change in % Correct from Null Model			7.8%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table I7

Logistic Regression Analysis Q2 UBS Total predicting Q2 BIMAS-2 Social Subscale Risk Status Membership (n = 205)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Total	-1.618	.312	26.922	1	< .001	.198
Constant	4.154	1.051	15.630	1	< .001	63.690
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			32.959	1	< .001	
Score test			33.429	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			5.897	6	.435	

Note. Cox & Snell $R^2 = .149$. Nagelkerke $R^2 = .237$. All statistics herein use three decimal places to maintain precision.

Table I8

Q2 UBS Total: Classification Table for Q2 BIMAS-2 Social Risk Status

Observed	Predicted		% Correct
	Yes (Concern)	No (No Concern)	
Yes (Concern)	11	29	27.50%
No (No Concern)	7	158	95.76%
Overall % Correct			82.44%
Change in % Correct from Null Model			1.9%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table I9

Logistic Regression Analysis Q2 UBS Total predicting Q2 BIMAS-2 Academic Functioning Subscale Risk Status Membership (n = 205)

Predictor	β	$SE \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Total	-3.564	.547	42.406	1	< .001	.028
Constant	10.533	1.776	35.175	1	< .001	37545.226
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			92.183	1	< .001	
Score test			82.407	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			6.949	6	.326	

Note. Cox & Snell $R^2 = .362$. Nagelkerke $R^2 = .573$. All statistics herein use three decimal places to maintain precision.

Table I10

Q2 UBS Total: Classification Table for Q2 BIMAS-2 Academic Functioning Risk Status

Observed	Predicted		% Correct
	Yes (Concern)	No (No Concern)	
Yes (Concern)	23	18	56.10%
No (No Concern)	8	156	95.12%
Overall % Correct			87.32%
Change in % Correct from Null Model			7.3%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Appendix J

Aim 4 Logistic Regression and Classification Tables:

Q1 UBS Social/Emotional Engagement Subscale and Q4 BIMAS-2 Subscale

Table J1

Logistic Regression Analysis Q1 UBS Social/Emotional Engagement Predicting Q4 BIMAS-2 Conduct Subscale Risk Status Membership (n = 198)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Social/Emotional	-1.417	.315	20.286	1	< .001	.242
Constant	2.850	1.016	7.862	1	.005	17.283
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			24.761	1	< .001	
Score test			27.699	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			14.428	6	.025	

Note. Cox & Snell $R^2 = .118$. Nagelkerke $R^2 = .214$. All statistics herein use three decimal places to maintain precision.

Table J2

Q1 UBS Social/Emotional Engagement Subscale: Classification Table for Q4 BIMAS-2 Conduct Risk Status

Observed	Predicted		% Correct
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	5	22	18.52%
No (Low Risk)	2	169	98.83%
Overall % Correct			87.88%
Change in % Correct from Null Model			1.5%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table J3

Logistic Regression Analysis Q1 UBS Social/Emotional Engagement predicting Q4 BIMAS-2 Negative Affect Subscale Risk Status Membership (n = 198)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Social/Emotional	-1.602	.323	24.576	1	< .001	.201
Constant	3.726	1.054	12.502	1	< .001	41.507
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			32.988	1	< .001	
Score test			35.373	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			6.359	6	.384	

Note. Cox & Snell $R^2 = .153$. Nagelkerke $R^2 = .258$. All statistics herein use three decimal places to maintain precision.

Table J4

Q1 UBS Social/Emotional Engagement Subscale: Classification Table for Q4 BIMAS-2 Negative Affect Risk Status

Observed	Predicted		% Correct
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	7	26	21.21%
No (Low Risk)	3	162	98.18%
Overall % Correct			85.35%
Change in % Correct from Null Model			2.1%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table J5

Multiple Logistic Regression Analysis Q1 UBS Social/Emotional Engagement and Gender Predicting Q4 BIMAS-2 Cognitive/Attention Subscale Risk Status Membership (n = 198)

Predictor	β	$SE \beta$	Wald's χ^2	df	p	e^β (odds ratio)
Gender	-1.173	.420	7.781	1	.005	.310
UBS Social/Emotional	-1.443	.299	23.235	1	< .001	.236
Constant	5.322	1.215	19.175	1	< .001	204.831
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			39.334	2	< .001	
Score test			37.719	2	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			15.845	8	.045	

Note. Cox & Snell $R^2 = .180$. Nagelkerke $R^2 = .271$. All statistics herein use three decimal places to maintain precision. Results indicate that for every one-point increase in a student's UBS Social/Emotional Engagement subscale score at Q1 the odds of that student being more likely to score in the *any risk* range on the BIMAS-2 Cognitive/Attention subscale at Q4 decreases by 76.4% [i.e., $(1-0.236) \times 100$], holding gender constant. Additionally, results of the multiple logistic regression found that for males, the odds of being more likely to be rated by teachers in the *any risk* range on the BIMAS-2 Cognitive/Attention subscale at Q4 was 3.23 times [i.e., $1/0.310$] that of female students, holding UBS Social/Emotional Engagement subscale score constant. Simple logistic regression results with only UBS Social/Emotional Engagement score as predictor found similar odds ratio and pseudo R^2 results (i.e., Cox & Snell $R^2 = .143$; Nagelkerke $R^2 = .215$; odds ratio = 0.248).

Table J6

Q1 UBS Social/Emotional Engagement Subscale and Gender as Covariate: Classification Table for Q4 BIMAS-2 Cognitive/Attention Risk Status

Observed	Predicted		% Correct
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	11	36	23.40%
No (Low Risk)	9	142	94.04%
Overall % Correct			77.27%
Change in % Correct from Null Model			1.0%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table J7

Logistic Regression Analysis Q1 UBS Social/Emotional Engagement predicting Q4 BIMAS-2 Social Subscale Risk Status Membership (n = 198)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Social/Emotional	-1.126	.322	12.218	1	< .001	.324
Constant	1.352	1.027	1.731	1	.188	3.864
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			12.445	1	< .001	
Score test			14.723	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			7.633	6	.266	

Note. Cox & Snell $R^2 = .061$. Nagelkerke $R^2 = .137$. All statistics herein use three decimal places to maintain precision.

Table J8

Q1 UBS Social/Emotional Engagement Subscale: Classification Table for Q4 BIMAS-2 Social Risk Status

Observed	Predicted		% Correct
	Yes (Concern)	No (No Concern)	
Yes (Concern)	1	16	5.88%
No (No Concern)	3	178	98.34%
Overall % Correct			90.40%
Change in % Correct from Null Model			-1.0%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table J9

Logistic Regression Analysis Q1 UBS Social/Emotional Engagement predicting Q4 BIMAS-2 Academic Functioning Subscale Risk Status Membership (n = 198)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Social/Emotional	-1.257	.286	19.376	1	< .001	.284
Constant	2.699	.949	8.083	1	.004	14.865
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			23.034	1	< .001	
Score test			24.839	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			.383	6	.999	

Note. Cox & Snell $R^2 = .110$. Nagelkerke $R^2 = .181$. All statistics herein use three decimal places to maintain precision.

Table J10

Q1 UBS Social/Emotional Engagement Subscale: Classification Table for Q4 BIMAS-2 Academic Functioning Risk Status

Observed	Predicted		% Correct
	Yes (Concern)	No (No Concern)	
Yes (Concern)	5	30	14.29%
No (No Concern)	2	161	98.77%
Overall % Correct			83.84%
Change in % Correct from Null Model			1.5%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Appendix K

Aim 4 Logistic Regression and Classification Tables:

Q1 UBS Academic Readiness Subscale and Q4 BIMAS-2 Subscales

Table K1

Logistic Regression Analysis Q1 UBS Academic Readiness predicting Q4 BIMAS-2 Conduct Subscale Risk Status Membership (n = 198)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^{β} (odds ratio)
UBS Academic Readiness	-1.126	.278	16.416	1	< .001	.324
Constant	1.549	.810	3.662	1	.056	4.708
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			18.214	1	< .001	
Score test			19.011	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			15.544	6	.016	

Note. Cox & Snell $R^2 = .088$. Nagelkerke $R^2 = .160$. All statistics herein use three decimal places to maintain precision.

Table K2

Q1 UBS Academic Readiness Subscale: Classification Table for Q4 BIMAS-2 Conduct Risk

Status

Observed	Predicted		% Correct
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	2	25	7.41%
No (Low Risk)	1	170	99.42%
Overall % Correct			86.87%
Change in % Correct from Null Model			0.5%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table K3

Logistic Regression Analysis Q1 Academic Readiness predicting Q4 BIMAS-2 Negative Affect Subscale Risk Status Membership (n = 198)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Academic Readiness	-1.286	.273	22.217	1	< .001	.276
Constant	2.273	.798	8.110	1	.004	9.707
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			26.289	1	< .001	
Score test			26.891	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			10.540	6	.104	

Note. Cox & Snell $R^2 = .124$. Nagelkerke $R^2 = .209$. All statistics herein use three decimal places to maintain precision.

Table K4

Q1 UBS Academic Readiness Subscale: Classification Table for Q4 BIMAS-2 Negative Affect Risk Status

Observed	Predicted		%
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	6	27	18.18%
No (Low Risk)	3	162	98.18%
Overall % Correct			84.85%
Change in % Correct from Null Model			1.5%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table K5

Logistic Regression Analysis Q1 UBS Academic Readiness and Gender predicting Q4 BIMAS-2 Cognitive/Attention Subscale Risk Status Membership (n = 198)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
Gender	-.939	.432	4.731	1	.030	.391
UBS Academic Readiness	-1.666	.290	32.918	1	< .001	.189
Constant	5.169	1.051	24.184	1	< .001	175.811
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			54.094	2	< .001	
Score test			50.438	2	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			2.244	8	.973	

Note. Cox & Snell $R^2 = .239$. Nagelkerke $R^2 = .359$. All statistics herein use three decimal places to maintain precision. Results indicate that for every one-point increase in a student's UBS Academic Readiness score at Q1 the odds of that student being more likely to score in the *any risk* range on BIMAS-2 Cognitive/Attention at Q4 decreases by 81.1% [i.e., $(1-0.189) \times 100$], holding gender constant. Additionally, results of the multiple logistic regression found that for males, the odds of being more likely to be rated by teachers in the *any risk* range on the BIMAS-2 Cognitive/Attention subscale at Q4 was 2.56 times [i.e., $1/0.391$] that of female students, holding UBS Social/Emotional Engagement subscale score constant. Simple logistic regression results with only UBS Social/Emotional Engagement score as predictor found similar odds ratio and pseudo R^2 results (i.e., Cox & Snell $R^2 = .219$; Nagelkerke $R^2 = .330$; odds ratio = 0.183).

Table K6

Q1 UBS Academic Readiness Subscale and Gender as Covariate: Classification Table for Q4 BIMAS-2 Cognitive/Attention Risk Status

Observed	Predicted		% Correct
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	23	24	48.94%
No (Low Risk)	10	141	93.38%
Overall % Correct			82.83%
Change in % Correct from Null Model			6.5%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table K7

Logistic Regression Analysis Q1 UBS Academic Readiness predicting Q4 BIMAS-2 Social Subscale Risk Status Membership (n = 198)

Predictor	β	SE β	Wald's χ^2	df	p	e^{β} (odds ratio)
UBS Academic Readiness	-1.314	.339	15.029	1	< .001	.269
Constant	1.478	.930	2.526	1	.112	4.385
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			17.011	1	< .001	
Score test			18.176	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			2.146	6	.906	

Note. Cox & Snell $R^2 = .082$. Nagelkerke $R^2 = .186$. All statistics herein use three decimal places to maintain precision.

Table K8

Q1 UBS Academic Readiness Subscale: Classification Table for Q4 BIMAS-2 Social Risk Status

Observed	Predicted		% Correct
	Yes (Concern)	No (No Concern)	
Yes (Concern)	1	16	5.88%
No (No Concern)	1	180	99.45%
Overall % Correct			91.41%
Change in % Correct from Null Model			0.0%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table K9

Logistic Regression Analysis Q1 UBS Academic Readiness predicting Q4 BIMAS-2 Academic Functioning Subscale Risk Status Membership (n = 198)

Predictor	β	SE β	Wald's χ^2	df	p	e^β (odds ratio)
UBS Academic Readiness	-1.782	.314	32.161	1	< .001	.168
Constant	3.743	.893	17.553	1	< .001	42.214
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			45.266	1	< .001	
Score test			44.643	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			5.141	6	.526	

Note. Cox & Snell $R^2 = .204$. Nagelkerke $R^2 = .337$. All statistics herein use three decimal places to maintain precision.

Table K10

Q1 UBS Academic Readiness Subscale: Classification Table for Q4 BIMAS-2 Academic Functioning Risk Status

Observed	Predicted		% Correct
	Yes (Concern)	No (No Concern)	
Yes (Concern)	11	24	31.43%
No (No Concern)	8	155	95.09%
Overall % Correct			83.84%
Change in % Correct from Null Model			1.5%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Appendix L

Aim 4 Logistic Regression and Classification Tables:

Q1 UBS Total and Q4 BIMAS-2 Subscales

Table L1

Logistic Regression Analysis Q1 UBS Total predicting Q4 BIMAS-2 Conduct Subscale Risk Status Membership (n = 198)

Predictor	β	$SE\ \beta$	Wald's X^2	df	p	e^β (odds ratio)
UBS Total	-1.418	.318	19.859	1	< .001	.242
Constant	2.668	.983	7.374	1	.007	14.412
Test			X^2	df	p	
Overall model evaluations						
Likelihood ratio test			23.836	1	< .001	
Score test			25.877	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			19.190	7	.008	

Note. Cox & Snell $R^2 = .113$. Nagelkerke $R^2 = .207$. All statistics herein use three decimal places to maintain precision.

Table L2

Q1 UBS Total: Classification Table for Q4 BIMAS-2 Conduct Risk Status

Observed	Predicted		% Correct
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	4	23	14.81%
No (Low Risk)	1	170	99.42%
Overall % Correct			87.88%
Change in % Correct from Null Model			1.5%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table L3

Logistic Regression Analysis Q1 UBS Total predicting Q4 BIMAS-2 Negative Affect Subscale Risk Status Membership (n = 198)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
UBS Total	-1.627	.326	24.959	1	< .001	.196
Constant	3.586	1.013	12.540	1	< .001	36.084
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			33.004	1	< .001	
Score test			34.531	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			5.267	7	.627	

Note. Cox & Snell $R^2 = .154$. Nagelkerke $R^2 = .259$. All statistics herein use three decimal places to maintain precision.

Table L4

Q1 UBS Total: Classification Table for Q4 BIMAS-2 Negative Affect Risk Status

Observed	Predicted		% Correct
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	7	26	21.21%
No (Low Risk)	0	165	100.00%
Overall % Correct			86.87%
Change in % Correct from Null Model			3.6%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table L5

Logistic Regression Analysis Q1 UBS Total and Gender predicting Q4 BIMAS-2 Cognitive/Attention Subscale Risk Status Membership (n = 198)

Predictor	β	$SE\ \beta$	Wald's χ^2	df	p	e^β (odds ratio)
Gender	-1.101	.434	6.445	1	.011	.333
UBS Total	-1.773	.326	29.540	1	< .001	.170
Constant	6.068	1.240	23.954	1	< .001	432.020
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			50.264	2	< .001	
Score test			47.008	2	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			17.806	8	.023	

Note. Cox & Snell $R^2 = .224$. Nagelkerke $R^2 = .337$. All statistics herein use three decimal places to maintain precision. Results indicate that for every one-point increase in a student's UBS Academic Readiness score at Q1 the odds of that student being more likely to score in the *any risk* range on BIMAS-2 Cognitive/Attention at Q4 decreases by 83.0% [i.e., $(1-0.170) \times 100$], holding gender constant. Additionally, results of the multiple logistic regression found that for males, the odds of being more likely to be rated by teachers in the *any risk* range on the BIMAS-2 Cognitive/Attention subscale at Q4 was 3.00 times [i.e., $1/0.333$] that of female students, holding UBS Social/Emotional Engagement subscale score constant. Simple logistic regression results with only UBS Social/Emotional Engagement score as predictor found similar odds ratio and pseudo R^2 results (i.e., Cox & Snell $R^2 = .196$; Nagelkerke $R^2 = .294$; odds ratio = 0.173).

Table L6

Q1 UBS Total and Gender as Covariate: Classification Table for Q4 BIMAS-2 Cognitive/Attention Risk Status

Observed	Predicted		% Correct
	Yes (Any Risk)	No (Low Risk)	
Yes (Any Risk)	19	28	40.43%
No (Low Risk)	9	142	94.04%
Overall % Correct			81.31%
Change in % Correct from Null Model			5.0%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table L7

Logistic Regression Analysis Q1 UBS Total predicting Q4 BIMAS-2 Social Subscale Risk Status Membership (n = 198)

Predictor	β	$SE \beta$	Wald's χ^2	df	p	e^{β} (odds ratio)
UBS Total	-1.321	.347	14.485	1	< .001	.267
Constant	1.784	1.042	2.930	1	.087	5.953
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			15.750	1	< .001	
Score test			18.021	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			8.611	7	.282	

Note. Cox & Snell $R^2 = .076$. Nagelkerke $R^2 = .172$. All statistics herein use three decimal places to maintain precision.

Table L8*Q1 UBS Total: Classification Table for Q4 BIMAS-2 Social Risk Status*

Observed	Predicted		% Correct
	Yes (Concern)	No (No Concern)	
Yes (Concern)	1	16	5.88%
No (No Concern)	3	178	98.34%
Overall % Correct			90.40%
Change in % Correct from Null Model			-1.0%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

Table L9

Logistic Regression Analysis Q1 UBS Total predicting Q4 BIMAS-2 Academic Functioning Subscale Risk Status Membership (n = 198)

Predictor	β	$SE \beta$	Wald's χ^2	df	p	e^{β} (odds ratio)
UBS Total	-1.685	.328	26.404	1	< .001	.186
Constant	3.849	1.022	14.174	1	< .001	46.962
Test			χ^2	df	p	
Overall model evaluations						
Likelihood ratio test			35.770	1	< .001	
Score test			36.965	1	< .001	
Goodness-of-fit test						
Hosmer & Lemeshow			5.791	7	.564	

Note. Cox & Snell $R^2 = .165$. Nagelkerke $R^2 = .272$. All statistics herein use three decimal places to maintain precision.

Table L10*Q1 UBS Total: Classification Table for Q4 BIMAS-2 Academic Functioning Risk Status*

Observed	Predicted		% Correct
	Yes (Concern)	No (No Concern)	
Yes (Concern)	9	26	25.71%
No (No Concern)	3	160	98.16%
Overall % Correct			85.35%
Change in % Correct from Null Model			3.1%

Note. Classification table created from logistic regression analysis with the cutoff of 0.50.

References

- Albers, C. A., Glover, T. A., & Kratochwill, T. R. (2007). Introduction to the special issue: How can universal screening enhance educational and mental health outcomes? *Journal of School Psychology, 45*, 113–116. <https://doi.org/10.1016/j.jsp.2006.12.002>
- American Educational Research Association (AERA), American Psychological Association (APA), & National Center on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.
- AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Anthony, C. J., Elliott, S. N., DiPerna, J. C., & Lei, P. W. (2020). Initial Development and Validation of the Social Skills Improvement System—Social and Emotional Learning Brief Scales-Teacher Form. *Journal of Psychoeducational Assessment, 39*(2), 166-181. <https://doi.org/10.1177/0734282920953240>
- Bear, G. G., Whitcomb, S. A., Elias, M. J., & Blank, J. C. (2015). SEL and schoolwide positive behavioral interventions and supports. In J. A. Durlak, C. E. Domitrovich, R. P. Weissberg, & T. P. Gullotta (Eds.) *Handbook of social and emotional learning: Research and practice* (pp. 453 – 467). New York, NY: Guildford Press.
- Bishara, A. J. & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods, 17*(3), 399-417. <https://doi.org/10.1037/a0028087>
- Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità, *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*.

- Blake, J. J., Gregory, A., James, M., & Hasan, G. W. (2016, September). Early warning signs: Identifying opportunities to disrupt racial inequities in school discipline through data-based decision making. *School Psychology Forum*, 10(3), 289-306.
- Bradshaw, C. P., Bottiani, J. H., Osher, D., & Sugai, G. (2014). The integration of positive behavioral interventions and supports and social and emotional learning. In M. D. Weist, N. A. Lever, C. P. Bradshaw, & J. S. Owens (Eds.) *Handbook of school mental health: Research, training, practice, and policy, Second Edition* (pp. 1-14). New York, NY: Springer Science & Business Media.
- Bradshaw, C. P., Mitchell, M. M., & Leaf, P. J. (2010). Examining the effects of schoolwide positive behavioral interventions and supports on student outcomes: Results from a randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions*, 12(3), 133-148. <https://doi.org/10.1177/1098300709334798>
- Bradshaw, C. P., Waasdorp, T. E., & Leaf, P. J. (2012). Effects of school-wide positive behavioral interventions and supports on child behavior problems. *Pediatrics*, 130(5), e1136–e1145. doi:10.1542/peds.2012-0243
- Bradshaw, C. P., Waasdorp, T. E., & Leaf, P. J. (2015). Examining variation in the impact of school-wide positive behavioral interventions and supports: Findings from a randomized controlled effectiveness trial. *Journal of Educational Psychology*, 107(2), 546-557. <http://dx.doi.org.eres.library.manoa.hawaii.edu/10.1037/a0037630>
- Briggs-Gowan, M. J., Horwitz, S. M., Schwab-Stone, M. E., Leventhal, J. M., & Leaf, P. J. (2000). Mental health in pediatric settings: distribution of disorders and factors related to

- service use. *Journal of the American Academy of Child & Adolescent Psychiatry*, 39(7), 841-849. <https://doi.org/10.1097/00004583-200007000-00012>
- Bullis, M., & Yovanoff, P. (2006). Idle hands: Community employment experiences of formerly incarcerated youth. *Journal of Emotional and Behavioral Disorders*, 14(2), 71-85. <https://doi.org/10.1177/10634266060140020401>
- Carran, D. T., & Scott, K. G. (1992). Risk assessment in preschool children: Research implications for the early detection of educational handicaps. *Topics in Early Childhood Special Education*, 12(2), 196-211. <https://doi.org/10.1177/027112149201200205>
- Carter, A. S., Briggs-Gowan, M. J., & Davis, N. O. (2004). Assessment of young children's social-emotional development and psychopathology: Recent advances and recommendations for practice. *Journal of Child Psychology and Psychiatry*, 45(1), 109-134. <https://doi.org/10.1046/j.0021-9630.2003.00316.x>
- Cash, R. E., & Nealis, L. K. (2004). *Mental health in the schools: It's a matter of public policy*. National Association of School Psychologists Public Policy Institute, Washington, DC.
- Chorpita, B. F., & Donkervoet, C. (2005). Implementation of the Felix consent decree in Hawaii. In *Handbook of mental health services for children, adolescents, and families* (pp. 317-332). Springer, Boston, MA.
- Civil Rights Data Collection (CRDC). (n.d.). Hawaii Department of Education, Honolulu, HI (Survey Year: 2017): LEA Characteristics and Membership. Retrieved June 19, 2021, from <https://ocrdata.ed.gov/profile/9/district/29005/summary>
- Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic.

- Collaborative for Academic, Social, and Emotional Learning. (2015). *2015 CASEL guide: Effective social and emotional learning programs – Middle and high school edition*. Chicago, IL: Author.
- Collaborative for Social and Emotional Learning. (2018, October). Connecting schoolwide SEL with other school-based frameworks. Retrieved from:
https://schoolguide.casel.org/uploads/2019/01/SEL_MTSS-and-PBIS.pdf
- Committee for Children. (2011). *Second Step for Kindergarten through Grade 5*. Seattle, WA: Author.
- Comrey, L.A. & Lee, H.B. (1992). *A first course in factor analysis (2nd ed.)*. Hillside, NJ: Lawrence Erlbaum Associates.
- Conners, C. K. (2008). *Conners third edition (Conners 3)*. Los Angeles, CA: Western Psychological Services.
- Cook, C. R., Frye, M., Slemrod, T., Lyon, A. R., Renshaw, T. L., & Zhang, Y. (2015). An integrated approach to universal prevention: Independent and combined effects of PBIS and SEL on youths' mental health. *School Psychology Quarterly*, 30(2), 166-183.
<https://doi.org/10.1037/spq0000102>
- Corcoran, R. P., Cheung, A. C., Kim, E., & Xie, C. (2018). Effective universal school-based social and emotional learning programs for improving academic achievement: a systematic review and meta-analysis of 50 years of research. *Educational Research Review*, 25, 56-72. <https://doi.org/10.1016/j.edurev.2017.12.001>

- Costello, E. J., Mustillo, S., Erkanli, A., Keeler, G., & Angold, A. (2003). Prevalence and development of psychiatric disorders in childhood and adolescence. *Archives of general psychiatry*, 60(8), 837-844. <https://doi.org/10.1001/archpsyc.60.8.837>
- Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, 10(1), 7.
- David, E. J. R., Schroeder, T. M., & Fernandez, J. (2019). Internalized racism: A systematic review of the psychological literature on racism's most insidious consequence. *Journal of Social Issues*, 75(4), 1057-1086. <https://doi.org/10.1111/josi.12350>
- Denham, S. A. (2005). *Assessing social-emotional development in children from a longitudinal perspective for the National Children's Study: Social-emotional compendium of measures*. Fairfax, VA: George Mason University.
- Denham, S. A. (2015) Assessment of SEL in educational contexts. In J. A. Durlak, C. E. Domitrovich, R. P. Weissberg, & T. P. Gullotta (Eds.) *Handbook of social and emotional learning: Research and practice* (pp. 285 – 300). New York, NY: Guildford Press.
- Dever, B. V., Dowdy, E., & DiStefano, C. (2018). Examining the stability, accuracy, and predictive validity of behavioral–emotional screening scores across time to inform repeated screening procedures. *School Psychology Review*, 47(4), 360-371. <https://doi.org/10.17105/spr-2017-0092.v47-4>
- Dowdy, E., Doane, K., Eklund, K., & Dever, B. V. (2013). A comparison of teacher nomination and screening to identify behavioral and emotional risk within a sample of

- underrepresented students. *Journal of Emotional and Behavioral Disorders*, 21(2), 127-137. <https://doi.org/10.1177/1063426611417627>
- 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(4), 166-176. <https://doi.org/10.1007/s12310-010-9036-3>
- Downey, D. B., & Pribesh, S. (2004). When race matters: Teachers' evaluations of students' classroom behavior. *Sociology of Education*, 77(4), 267-282. <https://doi.org/10.1177/003804070407700401>
- Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child development*, 82(1), 405-432. <https://doi.org/10.1111/j.1467-8624.2010.01564.x>
- Dvorsky, M. R., Girio-Herrera, E., & Owens, J. S. (2014). School-based screening for mental health in early childhood. In M. D. Weist et al. (Eds.) *Handbook of school mental health: Research, training, practice, and policy* (pp. 297-310). Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-7624-5_22
- Eisenberg, N., Fabes, R. A., Shepard, S. A., Murphy, B. C., Jones, S., & Guthrie, I. K. (1998). Contemporaneous and longitudinal prediction of children's sympathy from dispositional regulation and emotionality. *Child Development*, 69(3), 767-790. <https://doi.org/10.1111/j.1467-8624.1998.tb06242.x>
- Elias, M. J. (2006). The connection between academic and social-emotional learning. In M. J. Elias & H. Arnold (Eds.), *The educators guide to emotional intelligence and academic*

- achievement: Social-emotional learning in the classroom* (pp. 4-14). Thousand Oaks, CA: Corwin Press.
- Elliott, S. N., Davies, M. D., Frey, J. R., Gresham, F., & Cooper, G. (2018). Development and initial validation of a social emotional learning assessment for universal screening. *Journal of Applied Developmental Psychology, 55*, 39-51.
<https://doi.org/10.1016/j.appdev.2017.06.002>
- Elliott, S. N., Frey, J. R., & Davies, M. (2015). Systems for assessing and improving students' social skills to achieve academic competence. In J. A. Durlak, C. E. Domitrovich, R. P. Weissberg, & T. P. Gullotta (Eds.) *Handbook of social and emotional learning: Research and practice* (pp. 301 – 319). New York, NY: Guildford Press.
- Epstein, M. H., & Sharma, J. M. (1998). *Behavioral and Emotional Rating Scale: Examiner's manual*. Austin, TX: PRO-ED.
- Epstein, M. H., Mooney, P., Ryser, G., & Pierce, C. D. (2004). Validity and reliability of the Behavioral and Emotional Rating scale (2nd ed.): Youth Rating scale. *Research on Social Work Practice, 14*(5), 358–367. <https://doi.org/10.1177/1049731504265832>
- ESSA (2015). Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016).
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods, 4*(3), 272-299. <https://doi.org/10.1037/1082-989x.4.3.272>

- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(1), 1149-1160. <https://doi.org/10.3758/brm.41.4.1149>
- Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Thousand Oaks, CA: SAGE Publications.
- Fisher, R. A. (1921). On the “probable error” of a coefficient of correlation deduced from a small sample. *Metron* 1(4), 3–32. Reprinted in *Collected Papers of R. A. Fisher, Volume I: 1912–1924*, Ed. J. H. Bennett, 205–235. Adelaide: University of Adelaide, 1971. Text accessible at <http://digital.library.adelaide.edu.au/coll/special/fisher/14.pdf>.
- Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7(3), 286-299. <https://doi.org/10.1037/1040-3590.7.3.286>
- George, D., & Mallery, P. (2003). *Reliability analysis. SPSS for Windows, step by step: a simple guide and reference*. Boston: Allyn & Bacon.
- Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology*, 45(2), 117-135. <https://doi.org/10.1016/j.jsp.2006.05.005>
- Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, 38(5), 581–586. <https://doi.org/10.1111/j.1469-7610.1997.tb01545.x>
- Greenbaum, P. E., Dedrick, R. F., Friedman, R. M., Kutash, K., Brown, E. C., Lardieri, S. P., & Pugh, A. M. (1996). National Adolescent and Child Treatment Study (NACTS) outcomes

- for children with serious emotional and behavioral disturbance. *Journal of Emotional and Behavioral Disorders*, 4(3), 130-146. <https://doi.org/10.1177/106342669600400301>
- Greene, R. W., Beszterczey, S. K., Katzenstein, T., Park, K., & Goring, J. (2002). Are students with ADHD more stressful to teach? Patterns of teacher stress in an elementary school sample. *Journal of Emotional and Behavioral Disorders*, 10(2), 79-89. <https://doi.org/10.1177/10634266020100020201>
- Gresham, F. M. (2015). Evidence-based social skills interventions for students at risk for EBD. *Remedial and Special Education*, 36(2), 100-104. <https://doi.org/10.1177/0741932514556183>
- Gresham, F. M., & Elliott, S. N. (1990). *Social Skills Rating System*. Minneapolis, MN: Pearson Assessments.
- Gresham, F. M., Sugai, G., Horner, R. H., Quinn, M. M., & McInerney, M. (1998). *Classroom and schoolwide practices that support students' social competence: A synthesis of research*. Washington, DC: US Department of Education.
- Guadagnoli, E., & Velicer, W. F. (1988). Relation to sample size to the stability of component patterns. *Psychological Bulletin*, 103(2), 265-275. <https://doi.org/10.1037/0033-2909.103.2.265>
- Hackel, M., & Kan Hui, W. (2017, September). *The Behavior Intervention Monitoring Assessment System 2 (BIMAS-2TM)* [PowerPoint slides]. Office of Curriculum, Instruction, & Student Support (OCISS), <https://www.hawaiipublicschools.org>

Hawai'i State Department of Education (HIDOE). (December 13, 2018). *Strive Hawaii Report:*

Hale 'iwa Elementary 2017-2018. Retrieved from HIDOE website:

<http://www.hawaiipublicschools.org/Reports>

Hawai'i Department of Education (HIDOE) and Board of Education (BOE). (2016, December).

Strategic Plan 2017-2020 Executive Summary. Honolulu, HI: Author. Retrieved from:

<http://www.hawaiipublicschools.org>

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research:

Common errors and some comment on improved practice. *Educational and*

Psychological measurement, 66(3), 393-416. <https://doi.org/10.1177/0013164405282485>

Horner, R. H., Sugai, G., Smolkowski, K., Eber, L., Nakasato, J., Todd, A. W., & Esperanza, J.

(2009). A randomized, wait-list controlled effectiveness trial assessing school-wide

positive behavior support in elementary schools. *Journal of Positive Behavior*

Interventions, 11(3), 133-144. <https://doi.org/10.1177/1098300709332067>

Horwitz, S. M., Leaf, P. J., Leventhal, J. M., Forsyth, B., & Speechley, K. N. (1992).

Identification and management of psychosocial and developmental problems in

community-based, primary care pediatric practices. *Pediatrics*, 89(3), 480-485.

Hosmer, D. W., Jr., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York:

Wiley. <https://doi.org/10.1002/0471722146>

Humphrey, N., Kalamouka, A., Wigelsworth, M., Lendrum, A., Deighton, J., & Wolpert, M.

(2011). Measures of social and emotional skills for children and young people: A

systematic review. *Educational and Psychological Measurement*, 71(4), 617-637.

<https://doi.org/10.1177/0013164410382896>

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

Jenkins, L. N., Demaray, M. K., Wren, N. S., Secord, S. M., Lyell, K. M., Magers, A. M., ...

Tennant, J. (2014). A critical review of five commonly used social-emotional and behavioral screeners for elementary or secondary schools. *Contemporary School Psychology*. <http://dx.doi.org/10.1007/s40688-014-0026-6>

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis, I: Artificial data.

Applied Statistics, 21(2), 160-173. <https://doi.org/10.2307/2346488>

Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer.

<https://doi.org/10.1007/978-1-4757-1904-8>

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1), 141-151.

<https://doi.org/10.1177/001316446002000116>

Kaiser, H. F. (1970). A second-generation little jiffy. *Psychometrika*, 35, 401-415.

<https://doi.org/10.1007/bf02291817>

Kaiser, H. F. & Rice, J. (1974). Little jiffy, mark 4. *Educational and Psychological*

Measurement, 34(1), 111-117. <https://doi.org/10.1177/001316447403400115>

Kamphaus, R. W., & Reynolds, C. R. (2015). *Behavior Assessment System for Children—Third Edition (BASC-3): Behavioral and Emotional Screening System (BESS)*. Bloomington, MN: Pearson

Kamphaus, R. W., Thorpe, J. S., Winsor, A. P., Kroncke, A. P., Dowdy, E. T., & VanDeventer, M. C. (2007). Development and predictive validity of a teacher screener for child

- behavioral and emotional problems at school. *Educational and Psychological Measurement*, 67, 342–356. <http://dx.doi.org/10.1177/00131644070670021001>
- Kawamoto, K. (2016, December 7). Book review: Rainbows in Me: Values of Aloha. The Hawai‘i Herald. Retrieved from <https://www.thehawaiiherald.com>
- Kelm, J. L., & McIntosh, K. (2012). Effects of school-wide positive behavior support on teacher self-efficacy. *Psychology in the Schools*, 49(2), 137-147. <https://doi.org/10.1002/pits.20624>
- King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis* 9(2), 137–163. <https://doi.org/10.1093/oxfordjournals.pan.a004868>
- Kline, P. (1979). *Psychometrics and psychology*. London, UK: Academic Press.
- Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher*, 35(7), 3-12. <https://doi.org/10.3102/0013189x035007003>
- Lane, K. L., Oakes, W. P., Common, E. A., Brunsting, N., Zorigian, K., Hicks, T., & Lane, N. A. (2019). A comparison between SRSS-IE and BASC-2 BESS scores at the middle school level. *Behavioral Disorders*, 44(3), 162-174. <https://doi.org/10.1177/0198742918794843>
- Lane, K. L., Buckman, M. M., Oakes, W. P., & Menzies, H. M. (2020). Tiered systems and inclusion: Potential benefits, clarifications, and considerations. In K. L., Lane, H. M. Menzies, W. P. Oakes, & J. R. Kalberg (Eds.) *Developing a schoolwide framework to prevent and manage learning and behavior problems*. Guilford Publications.
- Lane, K. L., Kalberg, J. R., & Menzies, H. M. (2009). *Developing schoolwide programs to prevent and manage problem behaviors*. New York, NY: Guilford Press.

- Lane, K. L., Menzies, H. M., Oakes, W. P., Lambert, W., Cox, M., & Hankins, K. (2012). A validation of the Student Risk Screening Scale for internalizing and externalizing behaviors: Patterns in rural and urban elementary schools. *Behavioral Disorders, 37*(4), 244-270. <https://doi.org/10.1177/019874291203700405>
- Lane, K. L., Oakes, W. P., Cantwell, E. D., Menzies, H. M., Schatschneider, C., Lambert, W., & Common, E. A. (2016). Psychometric evidence of SRSS-IE scores in middle and high schools. *Journal of Emotional and Behavioral Disorders, 25*(4), 233-245. <https://doi.org/10.1177/1063426616670862>
- Lane, K. L., Oakes, W. P., Harris, P. J., Menzies, H. M., Cox, M., & Lambert, W. (2012). Initial evidence for the reliability and validity of the Student Risk Screening Scale for internalizing and externalizing behaviors at the elementary level. *Behavioral Disorders, 37*(2), 99-122. <https://doi.org/10.1177/019874291203700204>
- 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*(3), 160-172. <https://doi.org/10.1177/1044207310379123>
- Lane, K. L., Oakes, W. P., Swogger, E. D., Schatschneider, C., Menzies, H. M., & Sanchez, J. (2015). Student risk screening scale for internalizing and externalizing behaviors: Preliminary cut scores to support data-informed decision making. *Behavioral Disorders, 40*(3), 159-170. <https://doi.org/10.17988/0198-7429-40.3.159>

- Lannie, A. L., Coddling, R. S., McDougal, J. L., & Meier, S. (2010, June). The Use of Change-Sensitive Measures to Assess School-Based Therapeutic Interventions: Linking Theory to Practice at the Tertiary Level. In *School Psychology Forum*, 4(2), 1-14.
- LeBuffe, P. A., Shapiro, V. B., & Robitaille, J. L. (2018). The Devereux Student Strengths Assessment (DESSA) comprehensive system: Screening, assessing, planning, and monitoring. *Journal of Applied Developmental Psychology*, 55, 62-70.
<https://doi.org/10.1016/j.appdev.2017.05.002>
- Levitt, J. M., Saka, N., Romanelli, L. H., & Hoagwood, K. (2007). Early identification of mental health problems in schools: The status of instrumentation. *Journal of School Psychology*, 45(2), 163-191. <https://doi.org/10.1016/j.jsp.2006.11.005>
- Lewis, T. J., & Sugai, G. (1999). Effective behavior support: A systems approach to proactive schoolwide management. *Focus on Exceptional Children*, 31(6), 1-24.
<https://doi.org/10.17161/fec.v31i6.6767>
- Liu, Y., & Zumbo, B. D. (2007). The impact of outliers on Cronbach's coefficient alpha estimate of reliability: Visual analogue scales. *Educational and Psychological Measurement*, 67(4), 620-634. <https://doi.org/10.1177/0013164406296976>
- Losen, D., Hodson, C., Keith, M. A., Morrison, K., & Belway, S. (2015). *Are we closing the discipline gap? The Center for Civil Rights remedies*. Los Angeles, CA: University of California. Retrieved from <https://eScholarship.org>
- Marandos, S. (2020). *Universal screeners: A multi-gated approach to intervention*. (Publication No. 27961153) [Doctoral dissertation, University of Massachusetts Lowell], ProQuest Dissertations and Theses Database (PQDT).

- Martin, C. R., & Savage-McGlynn, E. (2013). A ‘good practice’ guide for the reporting of design and analysis for psychometric evaluation. *Journal of Reproductive and Infant Psychology*, 31(5), 449-455. <https://doi.org/10.1080/02646838.2013.835036>
- Mashburn, A. J., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2006). Teacher and classroom characteristics associated with teachers’ ratings of prekindergartners’ relationships and behaviors. *Journal of Psychoeducational Assessment*, 24, 367–380.
<http://doi.org/10.1177/0734282906290594>
- Matayoshi, K. S. (2016, August). Implementation of new Behavioral Intervention Monitoring Assessment System and guidelines [Memorandum]. Honolulu, HI: State of Hawai‘i Department of Education, Office of the Superintendent.
- McConaughy, S. H., Volpe, R. J., Antshel, K. M., Gordon, M., & Eiraldi, R. B. (2011). Academic and social impairments of elementary school children with attention deficit hyperactivity disorder. *School Psychology Review*, 40(2), 200-225.
<https://doi.org/10.1080/02796015.2011.12087713>
- McCurdy, B. L., Mannella, M. C., & Eldridge, N. (2003). Positive behavior support in urban schools: Can we prevent the escalation of antisocial behavior? *Journal of Positive Behavior Interventions*, 5(3), 158-170. <https://doi.org/10.1177/10983007030050030501>
- McDougal, J. L., Bardos, A. N., & Meier, S. T. (2011). *Behavior Intervention Monitoring Assessment System technical manual*. Toronto: Multi-Health Systems.
- McElwain, N. L., Olson, S. L., & Volling, B. L. (2002). Concurrent and longitudinal associations among preschool boys' conflict management, disruptive behavior, and peer rejection.

- Early Education and Development*, 13(3), 245-264.
https://doi.org/10.1207/s15566935eed1303_1
- McIntosh, K., Bennett, J. L., & Price, K. (2011). Evaluation of social and academic effects of school-wide positive behaviour support in a Canadian school district. *Exceptionality Education International*, 21(1), 46-60. <https://doi.org/10.5206/eei.v21i1.7669>
- Meier, S. T. (1997). Nomothetic item selection rules for tests of psychological interventions. *Psychotherapy Research*, 7(4), 419-427.
<https://doi.org/10.1080/10503309712331332113>
- Meier, S. T. (1998). Evaluating change-based item selection rules. *Measurement and evaluation in counseling and development*, 31(1), 15-27.
<https://doi.org/10.1080/07481756.1998.12068947>
- Meier, S. T. (2000). Treatment sensitivity of the PE Form of the Social Skills Rating Scales: Implications for test construction procedures. *Measurement and Evaluation in Counseling and Development*, 33, 144-156.
<https://doi.org/10.1080/07481756.2000.12069006>
- Meier, S. T. (2004). Improving design sensitivity through intervention-sensitive measures. *American Journal of Evaluation*, 25(3), 321-334.
<https://doi.org/10.1177/109821400402500304>
- Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological bulletin*, 111(1), 172-175. <https://doi.org/10.1037/0033-2909.111.1.172>

- Miller, F. G., Cohen, D., Chafouleas, S. M., 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(2), 184-196. <https://doi.org/10.1037/spq0000085>
- Muris, P., Meesters, C., & van den Berg, F. (2003). The strengths and difficulties questionnaire (SDQ). *European child & adolescent psychiatry*, 12(1), 1-8. <https://doi.org/10.1007/s00787-003-0298-2>
- Naglieri, J. A., LeBuffe, P. A., & Shapiro, V. B. (2011/2014). *The Devereux Student Strengths Assessment – Mini (DESSA-Mini): Assessment, technical manual, and user's guide*. Charlotte, NC: Apperson.
- Naglieri, J. A., LeBuffe, P., & Shapiro, V. B. (2011). Universal screening for social-emotional competencies: A study of the reliability and validity of the DESSA-mini. *Psychology in the Schools*, 48(7), 660–671. <https://doi.org/10.1002/pits.20586>
- Nakasato, J. (2000). Data-based decision making in Hawaii's behavior support effort. *Journal of Positive Behavior Interventions*, 2(4), 247-251. <https://doi.org/10.1177/109830070000200413>
- Nattino, G., Pennell, M. L., & Lemeshow, S. (2020). Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test. *Biometrics*, 76(2), 549-560. <https://doi.org/10.1111/biom.13249>
- Nelson, J. R., Benner, G. J., Lane, K., & Smith, B. W. (2004). Academic achievement of K-12 students with emotional and behavioral disorders. *Exceptional children*, 71(1), 59-73. <https://doi.org/10.1177/001440290407100104>

- Nelson, J. R., Martella, R. M., & Marchand-Martella, N. (2002). Maximizing Student Learning: The Effects of a Comprehensive School-Based Program for Preventing Problem Behaviors. *Journal of Emotional & Behavioral Disorders*, 10(3), 136-148.
<https://doi.org/10.1177/10634266020100030201>
- Nickerson, A. B., & Fishman, C. (2009). Convergent and divergent validity of the Devereux Student Strengths Assessment. *School Psychology Quarterly*, 24(1), 48–59.
<https://doi.org/10.1037/a0015147>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd Ed.) New York: McGraw-Hill. Inc.
- O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers*, 32, 396-402. <https://doi.org/10.3758/bf03200807>
- Panayiotou, M., Humphrey, N., & Wigelsworth, M. (2019). An empirical basis for linking social and emotional learning to academic performance. *Contemporary Educational Psychology*, 56, 193-204. <https://doi.org/10.1016/j.cedpsych.2019.01.009>
- Pas, E. T., & Bradshaw, C. P. (2014). What affects teacher ratings of student behaviors? The potential influence of teachers' perceptions of the school environment and experiences. *Prevention Science*, 15(6), 940-950. <https://doi.org/10.1007/s11121-013-0432-4>
- Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12), 1373-1379. [https://doi.org/10.1016/s0895-4356\(96\)00236-3](https://doi.org/10.1016/s0895-4356(96)00236-3)

- Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 96(1), 3-14.
<https://doi.org/10.1080/00220670209598786>
- Phillips, B. M., & Lonigan, C. J. (2010). Child and informant influences on behavioral ratings of preschool children. *Psychology in the Schools*, 47(4), 374-390.
<https://doi.org/10.1002/pits.20476>
- President's New Freedom Commission on Mental Health (2003). Achieving the promise: Transforming mental health care in America. Final Report DHHS Pub., Vol. SMA-03-3832. Rockville, MD: Retrieved September 2, 2019, from
<https://govinfo.library.unt.edu/mentalhealthcommission/reports>
- Rimm-Kaufman, S. E., & Hulleman, C. S. (2015). SEL in elementary school settings: Identifying mechanisms that matter. In J. A. Durlak, C. E. Domitrovich, R. P. Weissberg, & T. P. Gullotta (Eds.) *Handbook of social and emotional learning: Research and practice* (pp. 151 – 166). New York, NY: Guildford Press.
- Rosen, P. J., Vaughn, A. J., Epstein, J. N., Hoza, B., Arnold, L. E., Hechtman, L., ... & Swanson, J. M. (2014). Social self-control, externalizing behavior, and peer liking among children with ADHD – CT: A mediation model. *Social Development*, 23(2), 288-305.
<https://doi.org/10.1111/sode.12046>
- Ross, S. W., & Horner, R. H. (2007). Teacher Outcomes of School-Wide Positive Behavior Support. *Teaching Exceptional Children Plus*, 3(6).

- Ross, S. W., Romer, N., & Horner, R. H. (2012). Teacher well-being and the implementation of school-wide positive behavior interventions and supports. *Journal of Positive Behavior Interventions*, 14(2), 118-128. <https://doi.org/10.1177/1098300711413820>
- Rubin, K. H., Coplan, R. J., & Bowker, J. C. (2009). Social withdrawal in childhood. *Annual Review of Psychology*, 60(1), 141-171. <https://doi.org/10.1146/annurev.psych.60.110707.163642>
- Ruscio, J. (2014). Rational/theoretical approach to test construction. *The Encyclopedia of Clinical Psychology*, 1-5. <https://doi.org/10.1002/9781118625392.wbecp454>
- Saft, E. W., & Pianta, R. C. (2001). Teachers' perceptions of their relationships with students: Effects of child age, gender, and ethnicity of teachers and children. *School Psychology Quarterly*, 16(2), 125-141. <https://doi.org/10.1521/scpq.16.2.125.18698>
- School Digger (n.d.). Haleiwa Elementary School: Student enrollment information. Retrieved February 16, 2021, from <https://www.schooldigger.com/go/HI/schools/0003000172/school.aspx?t=tbStudents#aDetail>
- Seeley, J. R., Severson, H. H., & Fixsen, A. A. M. (2014). *Empirically based targeted prevention approaches for addressing externalizing and internalizing behavior disorders within school contexts*. In H. M. Walker & F. M. Gresham (Eds.), *Handbook of evidence-based practices for emotional and behavioral disorders: Applications in schools* (p. 307–323). The Guilford Press.
- Severson, H. H., Walker, H. M., Hope-Doolittle, J., Kratochwill, T. R., & Gresham, F. M. (2007). Proactive, early screening to detect behaviorally at-risk students: Issues,

- approaches, emerging innovations, and professional practices. *Journal of School Psychology*, 45(2), 193-223. <https://doi.org/10.1016/j.jsp.2006.11.003>
- Shapiro, V. B., Kim, B. K., Robitaille, J. L., & LeBuffe, P. A. (2017). Protective factor screening for prevention practice: Sensitivity and specificity of the DESSA-Mini. *School Psychology Quarterly*, 32(4), 449-464. <https://doi.org/10.1037/spq0000181>
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4), 591-611. <https://doi.org/10.1093/biomet/52.3-4.591>
- Sheng, Y., & Sheng, Z. (2012). Is coefficient alpha robust to non-normal data? *Frontiers in Psychology*, 3(34), 1-13. <https://doi.org/10.3389/fpsyg.2012.00034>
- Siceloff, E. R., Bradley, W. J., & Flory, K. (2017). Universal Behavioral/Emotional Health Screening in Schools: Overview and Feasibility. *Report on Emotional & Behavioral Disorders in Youth*, 17(2), 32-38.
- Smith-Millman, M. K., Flaspohler, P. D., Maras, M. A., Splett, J. W., Warmbold, K., Dinnen, H., & Luebke, A. (2017). *Differences between teacher reports on universal risk assessments. Advances in School Mental Health Promotion*, 10(4), 235-249. <https://doi.org/10.1080/1754730X.2017.1333914>
- Splett, J. W., Smith-Millman, M., Raborn, A., Brann, K. L., Flaspohler, P. D., & Maras, M. A. (2018). Student, teacher, and classroom predictors of between-teacher variance of students' teacher-rated behavior. *School Psychology Quarterly*, 33(3), 460-468. <http://doi.org/10.1037/spq0000241>
- Streiner, D. L. (1994). Figuring out factors: The use and misuse of factor analysis. *Canadian Journal of Psychiatry*, 39, 135-140. <https://doi.org/10.1177/070674379403900303>

- Sugai, G., & Horner, R. (2006). A promising approach for expanding and sustaining the implementation of school-wide positive behavior support. *School Psychology Review*, 35, 245–259. <https://doi.org/10.1080/02796015.2006.12087989>
- Sugai, G., Sprague, J. R., Horner, R. H., & Walker, H. M. (2000). Preventing school violence: The use of office discipline referrals to assess and monitor school-wide discipline interventions. *Journal of Emotional and Behavioral Disorders*, 8(2), 94-101. <https://doi.org/10.1177/106342660000800205>
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics, sixth edition*. Needham Heights, MA: Allyn & Bacon.
- The Hawn Foundation. (2011). *The MindUP curriculum: Brain-focused strategies for learning and living*. New York: Scholastic.
- Trevethan, R. (2017). Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health*, 5(307), 1-7. <https://doi.org/10.3389/fpubh.2017.00307>
- U. S. Department of Education. (2019). What Works Clearinghouse. Retrieved September 2, 2019, from <http://ies.ed.gov/ncee/wwc>.
- Substance Abuse and Mental Health Services Administration (SAMHSA), U.S. Department of Health and Human Services. (2002). *SAMHSA model programs: Model prevention programs supporting academic achievement*. Washington, DC: Author.
- U.S. Public Health Service (2000). *Report of the surgeon general's conference on children's mental health: A national action agenda*. Washington, DC: Department of Health and Human Services.

- Valiente, C., Swanson, J., & Eisenberg, N. (2012). Linking students' emotions and academic achievement: When and why emotions matter. *Child development perspectives*, 6(2), 129-135. <https://doi.org/10.1111/j.1750-8606.2011.00192.x>
- van Luling, L. M. (2015). *Externalizing and internalizing problems: Does one trump the other? A comparison of teacher referral methods for determining student risk of social, emotional, & behavioral health needs*. (Publication No. 3736736) [Doctoral dissertation, William James College], ProQuest Dissertations and Theses Database (PQDT).
- Vile Junod, R. E., DuPaul, G. J., Jitendra, A. K., Volpe, R. J., & Cleary, K. S. (2006). Classroom observations of students with and without ADHD: Differences across types of engagement. *Journal of School Psychology*, 44(2), 87-104. <https://doi.org/10.1016/j.jsp.2005.12.004>
- Waasdorp, T. E., Bradshaw, C. P., & Leaf, P. J. (2012). The impact of schoolwide positive behavioral interventions and supports on bullying and peer rejection: A randomized controlled effectiveness trial. *Archives of Pediatrics & Adolescent Medicine*, 166(2), 149-156. <https://doi.org/10.1001/archpediatrics.2011.755>
- Wagner, M., & Davis, M. (2006). How are we preparing students with emotional disturbances for the transition to young adulthood? Findings from the National Longitudinal Transition Study—2. *Journal of Emotional and behavioral disorders*, 14(2), 86-98.
- Wagner, M., Kutash, K., Duchnowski, A. J., Epstein, M. H., & Sumi, W. C. (2005). The children and youth we serve: A national picture of the characteristics of students with emotional disturbances receiving special education. *Journal of emotional and behavioral disorders*, 13(2), 79-96. <https://doi.org/10.1177/10634266050130020201>

- Weist, M. D., Lever, N. A., Bradshaw, C. P., & Owens, J. S. (2014). Further advancing the field of school mental health. In M. D. Weist, N. A. Lever, C. P. Bradshaw, & J. S. Owens (Eds.) *Handbook of school mental health: Research, training, practice, and policy, Second Edition* (pp. 1-140). New York, NY: Springer Science & Business Media.
https://doi.org/10.1007/978-1-4614-7624-5_1
- Weissberg, R. P., Durlak, J. A., Domitrovich, C. E., & Gullotta, T. P. (2015). Social and emotional learning; Past, present, and future. In J. A. Durlak, C. E. Domitrovich, R. P. Weissberg, & T. P. Gullotta (Eds.) *Handbook of social and emotional learning: Research and practice* (pp. 3-19). New York, NY: Guildford Press.
- Wigelsworth, M., Humphrey, N., Kalambouka, A., & Lendrum, A. (2010). A review of key issues in the measurement of children's social and emotional skills. *Educational Psychology in Practice*, 26(2), 173-186. <https://doi.org/10.1080/02667361003768526>
- Wilson, D. B., Gottfredson, D. C., & Najaka, S. S. (2001). School-based prevention of problem behaviors: A meta-analysis. *Journal of Quantitative Criminology*, 17(3), 247-272.
<https://doi.org/10.1023/a:1011050217296>
- Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2), 79-94.
<https://doi.org/10.20982/tqmp.09.2.p079>
- Zigmond, N. (2006). Twenty-four months after high school: Paths taken by youth diagnosed with severe emotional and behavioral disorders. *Journal of Emotional and Behavioral Disorders*, 14(2), 99-107. <https://doi.org/10.1177/10634266060140020601>

Zwick, W. R. & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99(3), 432-442.

<https://doi.org/10.1037/0033-2909.99.3.432>

Zins, J. E. & Elias, M. J. (2006). Social and emotional learning. In G. G. Bear & K. M. Minke (Eds.), *Children's needs III: Development, prevention, and intervention* (pp. 1-13).

Bethesda, MD: National Association of School Psychologists.

Zins, J. E., Payton, J. W., Weissberg, R. P., & O'Brien, M. U. (2007). Social and emotional learning for successful school performance. In G. Matthews, M. Zeidner, & R. D. Roberts (Eds.), *Series in affective science. The science of emotional intelligence: Knowns and unknowns* (pp. 376-395). New York, NY, US: Oxford University Press.

<https://doi.org/10.1093/acprof:oso/9780195181890.003.0015>

