The Pennsylvania State University

The Graduate School

College of Education

USING ITEM RESPONSE THEORY TO IMPROVE THE EFFICIENT MEASUREMENT OF ACADEMIC COMPETENCE

A Dissertation in

School Psychology

by

Christopher J. Anthony

© 2016 Christopher J. Anthony

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

May 2016
The dissertation of Christopher J. Anthony was reviewed and approved* by the following:

James C. DiPerna  
Professor of Education (School Psychology)  
Dissertation Adviser, Chair of Committee

Robert L. Hale  
Professor Emeritus of Education (School Psychology)

Pui-Wa Lei  
Associate Professor of Education (Educational Psychology)

Hoi K. Suen  
Distinguished Professor of Education (Educational Psychology)

Barbara A. Schaefer  
Associate Professor of Education (Educational and School Psychology)  
Professor in Charge for Graduate Programs in School Psychology

* Signatures are on file in the Graduate School
Abstract

Academic competence is a broad construct encompassing skills, attitudes and behaviors promoting students’ success in schools (DiPerna & Elliott, 2000). Although there are various measures of academic competence (e.g., the Academic Competence Evaluation Scales – Teacher Form; ACES-TF; DiPerna & Elliott, 2000), few studies have examined the efficiency with which this construct is measured. Thus, the purpose of this project was to examine and improve the efficiency of measurement of academic competence. Two complementary studies were completed to achieve this goal. The first utilized polytomous Item Response Theory (IRT) techniques to examine the measurement efficiency of the ACES-TF and identify a set of maximally efficient items (SMEI) for each ACES-TF subscale. The second was a pilot study for a brief measure of academic competence based on the SMEIs with an independent sample. Results provide insights regarding the measurement efficiency of the ACES-TF, psychometric properties of a brief measure of academic competence, and improving measurement efficiency of behavior rating scales.

Keywords: Academic Enablers, Short Form Scale Development, Item Response Theory, Academic Competence Evaluation Scales
# Table of Contents

List of Tables .................................................................................................................. vii

List of Figures .................................................................................................................... viii

Acknowledgments ............................................................................................................. ix

Chapter 1. Introduction .................................................................................................... 1

Chapter 2. Examining the Measurement Efficiency of the ACES-TF .................................... 9

  Rationale and Purpose ..................................................................................................... 11

  Method .............................................................................................................................. 11

  Participants ....................................................................................................................... 11

  Measure ........................................................................................................................... 12

  Procedures ....................................................................................................................... 13

  Data analysis .................................................................................................................. 13

  Results .............................................................................................................................. 15

  Tests of Assumptions ..................................................................................................... 15

  Model Comparison ........................................................................................................ 16

  Item Analysis ................................................................................................................ 16

  Preliminary Reliability/Precision Analyses ..................................................................... 17

  Discussion ....................................................................................................................... 21

Chapter 3. Development and Evaluation of a Brief Measure of Academic Competence ....... 25

  Purpose and Hypotheses ............................................................................................... 26

  Purpose ........................................................................................................................... 26

  Hypothesis 1 .................................................................................................................... 26

  Hypothesis 2 .................................................................................................................... 27
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis 3</td>
<td>27</td>
</tr>
<tr>
<td>Hypothesis 4</td>
<td>27</td>
</tr>
<tr>
<td>Hypothesis 5</td>
<td>27</td>
</tr>
<tr>
<td>Method</td>
<td>27</td>
</tr>
<tr>
<td>Participants</td>
<td>27</td>
</tr>
<tr>
<td>Measures</td>
<td>28</td>
</tr>
<tr>
<td>Brief Measure of Academic Competence</td>
<td>28</td>
</tr>
<tr>
<td>Social Skills Improvement System – Teacher Rating Scales</td>
<td>28</td>
</tr>
<tr>
<td>STAR Reading and Math</td>
<td>29</td>
</tr>
<tr>
<td>ACES-TF Engagement and Motivation Full Form Subscales</td>
<td>30</td>
</tr>
<tr>
<td>Procedures</td>
<td>30</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>34</td>
</tr>
<tr>
<td>Results</td>
<td>32</td>
</tr>
<tr>
<td>Evidence for Content Validity</td>
<td>32</td>
</tr>
<tr>
<td>Reliability/Precision</td>
<td>33</td>
</tr>
<tr>
<td>Structural Validity</td>
<td>34</td>
</tr>
<tr>
<td>Data screening</td>
<td>34</td>
</tr>
<tr>
<td>Identification</td>
<td>35</td>
</tr>
<tr>
<td>Overall fit</td>
<td>36</td>
</tr>
<tr>
<td>Loadings</td>
<td>36</td>
</tr>
<tr>
<td>Convergent and Discriminant Validity Analyses</td>
<td>39</td>
</tr>
<tr>
<td>Social skills and problem behaviors</td>
<td>39</td>
</tr>
<tr>
<td>Academic achievement</td>
<td>40</td>
</tr>
</tbody>
</table>
List of Tables

Table 1. Demographic Characteristics of Student Participants by Gender.................................12

Table 2. GRM Model RMSEAs by ACES-TF and Set of Maximally Efficient Items (SMEIs) Subscales........................................................................................................................................17

Table 3. Demographic Characteristics of Participants from the Pilot Sample .......................28

Table 4. Standardized Factor Loadings from the Pilot Sample CFA .........................................38

Table 5. Interfactor Correlations from the Pilot Sample CFA .....................................................39

Table 6. Correlations between BMAC Scale and Subscale Scores and SSIS-TRS Scale and Subscale Scores for Primary Elementary Students .................................................................................43

Table 7. Correlations between BMAC Scale and Subscale Scores and SSIS-TRS Scale and Subscale Scores for Intermediate Elementary Students.................................................................44
List of Figures

Figure 1. Comparison of TIFs for ACES-TF and SMEI Academic Skills Subscales. ........19
Figure 2. Comparison of TIFs for ACES-TF and SMEI Academic Enablers Subscales. ....20
Figure 3. Comparison of TIFs for ACES-TF SMEI and BMAC Interpersonal Skills and Study Skills Subscales..................................................................................................................34
Figure 4. Pilot Sample Correlated Factors Confirmatory Factor Analysis Model ..........37
Acknowledgments

The completion of this dissertation and my graduate education would not have been possible without the support and guidance of many individuals.

First, I am indebted to my adviser, Dr. Jim DiPerna. I greatly appreciate the countless hours you have poured into my education. You have been a wonderful mentor, teacher, and friend. It has been a privilege to learn from you and I hope to emulate your thoughtful, dedicated approach to educational science in my future work.

I would also like to thank the members of my dissertation committee, Drs. Pui-Wa Lei, Robert Hale, and Hoi Suen. I have spent many hours in each of your respective offices asking questions and learning from you. Thank you for your guidance.

I must also thank the Institute of Education Sciences and specifically the Training Interdisciplinary Educational Scientists program for funding my graduate education and providing me with access to all the resources I needed as I progressed through my doctoral program. It has been a wonderful blessing to be a part of this program.

There are many other professors and teachers who have influenced me throughout my education. A few individuals who stand out include Dr. Darcia Narvaez, Dr. Dan Lapsley, Dr. Shirley Woika, and Dr. Paul Amato. I thank all of my teachers who have built on the work of each other at every juncture in my academic career.

I would also like to thank my cohortmates and friends who helped me through graduate school. Specifically, I thank Andrew, Josh, Ryan, Tom, Joe, Peter, Fr. Matt, Fr. Matthew, and Fr. Will for being wonderful friends throughout my time in State College.
I would not have completed this dissertation without the support of many family members. I must specifically offer my gratitude to the following individuals:

To my in-laws, thank you for accepting me into your family with open arms.

To my sister, thank you for your fierce loyalty and tireless support.

To my grandparents, thank you for your lectures and wisdom throughout the years.

To my mother, thank you for teaching me how crucial laughter is for a good life.

To my father, thank you for instilling in me a thirst for knowledge from an early age.

To my beautiful daughter Gianna, thank you for your zest for life. Your presence keeps me grounded and focused on what is truly important. It has been a joy to watch you grow and I can’t wait for the years to come.

To my devoted wife Lindsey, thank you for putting up with me throughout this process. I know it’s been a long road, but your patience and love have been instrumental in the finishing of this project and my graduate education. I love you so much and I’m blessed beyond measure to be your husband.

Finally, and most importantly, I thank God for providence in years past and ask for grace in years to come.
Promoting students’ academic achievement is one of the main goals of education systems throughout the world. As a result, researchers have examined numerous variables thought to influence academic achievement (e.g., Caprara, Barbaranelli, Pastorelli, Bandura, & Zimbardo, 2000; Singh, Granville, & Dika, 2002). Furthermore, broad models including several of these variables have shed light on factors promoting academic achievement and the relationships between these factors. Early models such as the one proposed by Carroll (1963) influenced later researchers (e.g., Bennett, 1978; Bloom, 1980; Cooley & Leinhardt, 1975) to develop various models of academic achievement. Haertel, Walberg, and Weinstein (1983) reviewed several of these models and noted that they all included a set of prerequisite characteristics (of either the student or the educational environment) influencing academic achievement, a set of instructional variables influencing academic achievement, and an outcome representing successful school learning or achievement. Despite these similarities, each model included different variables within these sets of common constructs.

Building on these early researchers, DiPerna (1999) incorporated the work of several other relevant researchers (Wentzel, 1993; Malecki, 1998) to develop a synthesized model of academic competence. The results of this synthesis informed the development of the term academic enablers, which refers to the “attitudes and behaviors that allow a student to participate in, and ultimately benefit from, academic instruction in the classroom” (DiPerna & Elliott, 2002). Specific academic enablers described include interpersonal skills, motivation, engagement, and study skills. These constructs have each been extensively researched individually (e.g., Peterson, 1990; Flook, Repetti, & Ullman, 2005) and were initially identified
through a variety of methods including a review of the literature, discussions with educational professionals, and empirical research (DiPerna & Elliott, 2002). Researchers (e.g., DiPerna, Volpe, & Elliott, 2002; 2005) have linked each of these variables with academic achievement and integrated them into a broad model of academic competence.

Interpersonal skills refer to “cooperative learning behaviors that allow students to interact effectively in academic settings” (DiPerna, 2004, p. 65). The social context of learning has been a major area of research throughout the history of educational psychology (e.g., Brofenbrenner, 1989; Vygotsky, 1978). Interpersonal skills can exert their influence on academic achievement in two ways (Wentzel & Watkins, 2002). First, interpersonal skills have been linked with motivational constructs (e.g., Flook et al., 2005) which further enable children to engage in and benefit from classroom instruction (DiPerna et al., 2002; 2005). Second, interpersonal skills can act directly by facilitating social classroom interactions that are primarily focused on academic tasks (Wentzel & Watkins, 2002). These interactions can directly facilitate academic achievement (e.g., Caprara et al., 2000; Malecki & Elliott, 2002). Much research has borne out this connection between interpersonal skills and academic achievement. For example, in a large meta-analysis of 213 studies, Durlak, Weissberg, Dymnicki, Taylor, and Schellinger (2011) found that elementary through high school aged children involved in social and emotional learning programs achieved an 11 percentile point gain in achievement as a result of participation in such programs. These results indicate that not only are interpersonal skills related to academic achievement, but also that improving these skills can improve children’s academic achievement. In these ways, interpersonal skills are a foundational and important academic enabler.

The construct of behavioral engagement includes behaviors such as writing, reading aloud or silently, discussing relevant topics, and asking or answering questions (Greenwood,
Horton, & Utley, 2002). Generally, if a child is actively or passively participating in classroom
instruction, he or she can be considered to be academically engaged. Engagement is a
multidimensional construct, and there are various different definitions of engagement that have
been used in research (Appleton, Christenson, & Furlong, 2008). Despite these complexities,
academic engagement has been found to be related to academic achievement and positive
socioemotional outcomes (Klem & Connell, 2004). For example, Finn (1993) found that even
within racial, linguistic, and socioeconomic status subgroups, engagement predicted academic
achievement. Also, academic engagement is related to both science and mathematics
achievement (Singh et al., 2002) and is one of the most important factors predicting school
dropout (Appleton et al., 2008). Academic engagement is also malleable and able to be
influenced by interventions (e.g., Christenson, Reschly, Appleton, Berman, Spanjers, & Varro,
2008; Greenwood et al., 1979; Hops & Cobb, 1973). These research findings indicate the
importance of engagement as an academic enabler that promotes student success in schools.

Motivation is defined as the characteristic, attribute, or attitude that moves individuals to
engage in a particular action (Gredler, 2001). This broad definition of motivation includes many
motivational constructs including attributional style (e.g., Borkowski, Weyhing, & Carr, 1988;
Gibb, Alloy, Walshaw, Corner, Shen, & Villari, 2006), self-efficacy (e.g., Zimmerman, Bandura,
& Martinez-Ponz, 1992), and intrinsic motivation (Cerasoli, Nicklin, & Ford, 2014). Most
contemporary models of motivation are social-cognitive and multifaceted. As such, they are not
concerned merely with whether a student is or is not motivated but also with the various
situational, dispositional, and content specific factors that affect a child’s motivation to complete
a task. In general, motivation is thought to be one of the most important factors promoting
academic achievement (Mitchell, 1992) and a large number of studies have supported this view
(e.g., Bandura, 1997; Peterson, 1990). For example, in a recent large scale meta-analysis of 154 studies, Cerasoli et al. (2014) found that in school based settings intrinsic motivation was found to be a small to medium sized ($\rho = .26$) predictor of performance. Thus, motivation has a critical role in broad models of academic achievement.

A student’s study skills include “behaviors or strategies that facilitate the processing of new material” (DiPerna, 2004, p. 65). Study skills have been extensively researched for decades (Richardson, Robnolt, & Rhodes, 2010). Themes identified by Richardson et al. (2010) that have emerged in the past few decades include a focus on motivation, the description of specific activities and programs, the specification of the construct of metacognition, the creation of assessments of study skills, and the utilization of study skills in increasingly advanced technological applications, among other developments (e.g., Falkenberg & Barbeta, 2013; Paulsen & Sayeski, 2013). Study skills are related to academic achievement (Gettinger & Seibert, 2002) and able to be influenced by specifically designed interventions (e.g., Falkenberg & Barbeta, 2013).

Despite the research base supporting the DiPerna model (e.g., Anthony, DiPerna, & Amato, 2014; DiPerna et al., 2002; 2005), the academic enablers identified in the model are not an exhaustive list of potential non-academic student variables influencing academic achievement. Other researchers (Duckworth, Peterson, Matthews, & Kelly, 2007; Li-Grining, Votruba-Drzal, Maldonado-Carreno, & Haas, 2010; McDermott, 1984) have demonstrated similar links between constructs such as grit (Duckworth, 2006), learning behaviors (McDermott, 1984; McDermott & Beitman, 1984), and approaches to learning (Kagan, Moore, & Bredenkamp, 1995; U.S. Department of Education, 1996) and academic achievement. These
constructs and the growing interest in them further indicate the importance of non-academic skills and behaviors that promote academic achievement.

In addition to academic enablers, DiPerna’s (e.g., 1999) model includes a student’s current level of fundamental academic skills. Academic skills are broadly defined as the basic and complex skills that are taught in school (DiPerna & Elliott, 2000). Academic skills are the major focus of most curricula and, in conjunction with academic enablers and other constructs (e.g., executive functions) influence academic achievement. Although there are many different domains of academic skills, DiPerna included reading and language arts, mathematics, and critical thinking in his model of academic competence. Including academic skills in a broader model of academic competence emphasizes the relationship between academic enablers and academic skills and provides opportunities to better isolate areas of student need.

Adequate and technically sound measurement is crucial for the development and evaluation of any construct. Furthermore, for any measure to be useful, it must be well adapted to the context in which it is intended to be used. Recent developments in the American educational context have led to increased importance of measurement efficiency and brevity in socio-emotional and behavioral measures. Specifically, multi-tiered service delivery systems have been developed for many important outcomes (e.g., School Wide Positive Behavior Intervention and Support; Sugai & Horner, 2009). These multi-tiered systems are marked by various levels of service delivery with increasingly focused and involved assessment and intervention as children progress to higher tiers of supports. Although there are many advantages to such systems, they also put demands on teachers’ time and require measures that are specifically adapted to their purpose within the system (Sugai & Horner, 2009).
Measurement efficiency and brevity are especially important for measures intended to be used at the primary and secondary tiers of these multi-tiered models (Malecki & Demaray, 2007).

Similarly, educational researchers are often faced with limited resources and, as a result, face challenges when utilizing long rating scales developed for more intensive purposes. Such challenges include a greater amount of missing data (Stanton, Sinar, Balzer, & Smith, 2002) and increased error due to fatigue or boredom (Galesic & Bosnjak, 2009). Limitations arising from long measures often necessitate the development of short forms. For example, citing requests from personality researchers, Rammstedt and John (2007) shortened the already brief 44 item version of the Big Five Inventory (BFI-44; John, Donahue, & Kentle, 1991) to 10 items that could be completed in 1 minute or less. Although this study administered a shortened form independently, researchers often assume psychometric adequacy without administering a developed shortened form independently (Smith, McCarthy, & Anderson, 2000). These and other common questionable practices plague short form development (Smith, et al., 2000) and highlight the need for well-developed and independently validated short forms of common measures. Ultimately, while there are numerous measures currently available for use, there are indications that many may not be well adapted to various applications in the current educational and research climate in which time and resources are significantly limited.

There are several published measures of constructs similar to academic enablers (e.g., Learning Behaviors Scale; LBS; McDermott, Green, Francis, & Stott, 1999). Furthermore, there are many measures of individual academic enablers (e.g., School Motivation and Learning Strategies Inventory; Stroud & Reynolds, 2006) and direct measures of academic skills (e.g., the Woodcock Johnson Tests of Achievement – Fourth Edition; Schrank, Mather, & McGrew, 2014). Perhaps the most widely used (Cleary et al., 2010) measure that assesses both academic
skills and enablers is the Academic Competence Evaluation Scales – Teacher Form (ACES-TF; DiPerna & Elliott, 2000), which includes 73 items measuring two broad domains: academic skills and academic enablers. The Academic Skills scale assesses the domains of Reading/Language Arts, Mathematics, and Critical Thinking. The Academic Enablers scale measures competence in the domains of Interpersonal Skills, Motivation, Engagement, and Study Skills. The ACES-TF was standardized with 1,000 students. For this sample, Cronbach’s α coefficients for scores from the ACES-TF ranged from .94 to .99, and two three week interval stability coefficients ranged from .88 to .97 across subscales and age ranges (DiPerna & Elliott, 2000).

Furthermore, scores from the ACES-TF have shown evidence of convergent and discriminant validity. For example, scores were moderately to highly correlated ($r = .40 - .71$ across subtests) with scores from the Social Skills domain of the Social Skills Rating System (SSRS; Gresham & Elliott, 1990; DiPerna & Elliott, 2000). Also, scores from the ACES-TF were negatively correlated ($r = -.25$ to -.69) with scores from the Problem Behaviors domain of the SSRS (DiPerna & Elliott, 2000). There is also evidence for the structural validity of the ACES-TF in the form of exploratory factor analyses conducted with school-aged (DiPerna & Elliott, 1999; 2000) and college-aged (DiPerna, 2004) samples. The ACES-TF has also been employed in various research studies examining topics such as attention deficit/hyperactivity disorder (Volpe, DuPaul, DiPerna, Jitendra, Lutz, Tresco & Junod, 2006; Demaray & Jenkins, 2011), intervention evaluation (Oakes, Lane, Cox, Magrane, Jenkins, Hankins, 2012), cross cultural research (Grigorenko, Kornev, Rakhlin, & Krivulskaya, 2011) and teacher-child relationships (McCormick, O’Connor, Cappella, & McClowery 2013).

Despite the strengths of the ACES-TF, it has limitations as well. Specifically, the ACES-TF includes 73 items, which may limit the ability of educational practitioners to examine ACES-
TF constructs in certain contexts (e.g., at the secondary tier within a multi-tiered framework). The ACES-TF length may also be prohibitive for many research studies examining ACES-TF constructs. Thus, this project included two studies designed to address two complementary goals: examining and improving the efficiency of measurement of academic competence. In the first study, Item Response Theory (IRT) methodology was used to analyze the items of the ACES-TF and identify a set of maximally efficient items (SMEI) for each ACES-TF subscale. In the second study, a brief measure of academic competence (BMAC) informed by the SMEIs identified in the first study was piloted. As the ACES-TF was developed within a Classical Test Theory (CTT) framework, conducting IRT analyses holds promise for improving the measurement efficiency of the ACES-TF (e.g., Anthony, DiPerna, & Lei, 2016). Ultimately, the results of this dissertation address gaps in the research base for the ACES-TF and extend that research base by developing a BMAC.
Chapter 2

Examining the Measurement Efficiency of the ACES-TF

One of the most important practical aspects of rating scales is the efficiency with which constructs are measured. This is especially true given the wide variety of contexts within which rating scales are used. In addition to traditional psychoeducational evaluation contexts, rating scales have become commonplace in contexts such as universal screening (e.g., the Behavioral and Emotional Screening System; Kamphaus & Reynolds, 2007) and large scale educational research (e.g., the Early Childhood Longitudinal Study – Kindergarten Cohort Class of 1998-99). The strengths of many published rating scales (e.g., comprehensive construct coverage, high reliability coefficients, etc.) are not well matched to the needs of these contexts (e.g., brevity). Teacher rating scales are particularly susceptible to these difficulties, given the increasing time pressures on teachers (Nias, 1989; Cockburn, 1994; Skaalvik & Skaalvik, 2010). Thus, measurement efficiency is a particularly important consideration for teacher rating scale development.

There are two prominent measurement frameworks that have been used for developing rating scales – Classical Test Theory (CTT) and Item Response Theory (IRT; Goetz et al., 2013). Scale development within a CTT model has several important advantages including “weak” assumptions (i.e., assumptions that are met by most data sets; Hambleton & Swaminathan, 1985) and relatively low sample size requirements (Zickar & Broadfoot, 2009). Furthermore, CTT has been utilized extensively in the development of most psychological measures throughout the past century. Despite its widespread use, one major weakness of CTT in the context of measurement efficiency is the dependence of item statistics on the sample (of items and of participants) that generated them. For example, item discrimination statistics ($p$) tend to be higher in
heterogeneous samples than in homogenous samples (Hambleton & Jones, 1993). This sample
dependence can limit the ability of test developers to design flexible tests adapted to suit certain
contexts (and thus more efficient for that context).

IRT is a measurement framework in which a number of parameters for each item and
respondent are estimated (typically with maximum likelihood procedures) and analyzed.
Methodology based on IRT has several advantages for maximizing measurement efficiency
(Edelen & Reeve, 2006). For example, when relevant IRT assumptions are met (e.g., local
independence, unidimensionality) and the employed model fits the data well, IRT item
parameters are sample independent (Hambleton, Swaminathan, & Rogers, 1991). Sample
independent item parameters allow test developers to consider and modify item functioning
rather than considering entire scales or tests as a whole.

Another advantage of IRT for maximizing measurement efficiency is its ability to
provide information (akin to precision or reliability) values of an item across levels of the latent
ability or trait (often referred to as the θ parameter) being measured. This advantage allows test
developers to prioritize and select items based on the ability level at which they provide the most
information (Hambleton & Jones, 1993). Tests constructed using IRT-informed approaches can
be designed to more efficiently measure a construct at a particular range of ability (e.g., at-risk
students) corresponding to a particular intended purpose of measurement (e.g., universal
screening).

In addition to its advantages, IRT has several limitations as well. Most importantly, the
assumptions underlying IRT are more difficult to meet than those underlying CTT (Hambleton &
Swaminathan, 1985). If these strong assumptions are not met (particularly the assumption of
local independence), resulting parameters, information functions, and item characteristic curves
are biased and cannot be interpreted (Reeve & Fayers, 2005). Nonetheless, IRT methods have become more prominent in studies examining the measurement efficiency of behavior rating scales (e.g., Anthony et al., 2016).

**Rationale and Purpose**

As noted in Chapter 1, the ACES-TF (DiPerna & Elliott, 2000) is a commonly used teacher rating scale to assess academic competence. Because the ACES-TF was developed within a CTT framework, it may have been susceptible to some of the aforementioned limitations associated with CTT. DiPerna and Elliott (2000, p. 108) briefly reference measurement efficiency in the ACES-TF manual stating, “Anecdotal evidence from teachers suggests the ACES is time-efficient.” Beyond this reference, however, no systematic examination of the measurement efficiency of the ACES-TF has been conducted. As such, the purpose of this study was to examine the measurement efficiency of the ACES-TF using polytomous IRT and identify a set of maximally efficient items (SMEI) for each subscale of the ACES-TF.

**Method**

**Participants.** The standardization sample for the ACES-TF (DiPerna & Elliott, 2000) was used to identify subscale SMEIs. Children from the first through fifth grade were included in this study. As a result, the final sample included 455 students. Demographic characteristics of these students are presented in Table 1.
Table 1. *Demographic Characteristics of Student Participants by Gender (N = 455)*

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Girls (n = 221)</th>
<th>Boys (n = 234)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>63</td>
<td>63</td>
</tr>
<tr>
<td>Minority</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Hispanic</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>Other</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Grade</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>Second</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>Third</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Fourth</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>Fifth</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>Region</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northeast</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td>Northcentral</td>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>South</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>West</td>
<td>28</td>
<td>24</td>
</tr>
<tr>
<td>Educational Status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General Education</td>
<td>77</td>
<td>71</td>
</tr>
<tr>
<td>Special Education</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>At Risk</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>SES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>29</td>
<td>24</td>
</tr>
<tr>
<td>Low-Middle</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>Middle</td>
<td>24</td>
<td>32</td>
</tr>
<tr>
<td>Middle-High</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>High</td>
<td>9</td>
<td>7</td>
</tr>
</tbody>
</table>

*Note.* Percentages do not sum to 100 in some cases due to missing data and/or rounding.

**Measure.** The Academic Competence Evaluation Scales – Teacher Form (ACES-TF; DiPerna & Elliott, 2000) was used for this project. The ACES-TF is a comprehensive, 73 item teacher rating scale of the academic skills and academic enablers of students in kindergarten through the twelfth grade. Items are rated on a 5-point Likert scale from 1 (*Far Below*) to 5 (*Far Above*) for Academic Skills scales and subscales (in reference to grade level expectations) and from 1 (*Never*) to 5 (*Almost Always*) for Academic Enablers scales and subscales. Academic
Skills subscales include Reading/Language Arts, Mathematics, and Critical Thinking. Academic Enablers subscales include Interpersonal Skills, Engagement, Motivation, and Study Skills. There is evidence for the reliability and validity of the ACES-TF in the form of acceptable Cronbach’s α and stability coefficients, expected relationships with concurrent measures of similar skills, and factor analytic evidence supporting the proposed structure of the measure (DiPerna & Elliott, 2000).

**Procedures.** Teachers provided ratings for up to six children in their classroom. Teachers were further instructed to select an equal number of students who were in general education classrooms, who had been identified as having a learning disability, and who were at risk for academic failure. Finally, they were instructed to rate an equal number of boys and girls and include 1 to 2 students of ethnic minority in their ratings.

**Data analysis.** As mentioned above, IRT analyses were used to examine the measurement efficiency of the ACES-TF. Because item response data were ordinal, polytomous IRT methods were used. Relevant assumptions of IRT (e.g., essential unidimensionality, local independence) were checked by subscale as part of these analyses. Subsequently, several polytomous IRT models were fit to the data and compared. Specifically, the following polytomous IRT models were tested with each subscale: the Partial Credit Model (PCM; Masters, 1982), the Generalized Partial Credit Model (GPCM; Muraki, 1992), and the Graded Response Model (GRM; Samejima, 1970). Once a superior fitting model was identified, item parameters were computed with this model and used to generate item information curves (IICs). This information as well as content consideration was all utilized in the identification of SMEI for each ACES-TF subscale. These analyses were conducted at the subscale level due to the

---

1 For teachers rating children in Grades 1 and 2, instructions were modified to include equal numbers of students in the bottom, middle, and upper third of their class.
theoretical distinctions between academic skills and enablers subconstructs and to preserve the
breadth of measurement of the original ACES-TF.

Missing data were addressed in several ways. First, items that were specifically
designated to be applicable to only students in the third through twelfth grades were eliminated.
Next, because all IRT analyses were completed at the subscale level, missing data were analyzed
at the subscale level as well. Cases missing more than 50% of their values were deleted by
subscale for all analyses (e.g., A case was deleted if it was missing over 50% of its values for one
scale for all analyses of that scale but that same case was included in analyses of scales for which
that case was not missing over 50% of its values). After this subscale specific listwise deletion,
there were still several missing values on many items. The percentages of missing values for
items after implementation of subscale-specific listwise deletion ranged from 0 to 17.5 (median =
0.4).

Little’s (1988) test for Missing Completely at Random (MCAR) was conducted by
subscale to determine if any systematic missing data patterns were present. All $\chi^2$ values were
not statistically significant ($p = .05$) except for the Mathematics, Engagement, and Study Skills
subscales. Of these scales, only 3 items were missing 5% or more of their cases. Two of these
items made explicit reference to tests (Item 45 and Item 66) and one made reference to complex
problem solving. As expected, most of these missing cases (54% for Item 18, 73% for Item 45,
and 55% for Item 66) were for first graders. These findings were taken to indicate that cases
were missing because they referenced activities (e.g., tests) that were not applicable for some
students based on grade and classroom (observed variables) and thus Missing at Random (MAR;
Pigot, 2001).
For these missing cases, two different missing data procedures were used. First, because IRT overall fit and local dependence statistics are particularly sensitive to matrix sparsity, the relative mean substitution (RMS; Raaijmakers, 1999) method of imputation was used. This form of imputation was developed specifically for Likert items and estimates the relative position of a respondent in a sample by calculating the ratio of the intra-individual mean of all nonmissing item scores to the grand mean of all valid item scores. This relative position is then multiplied by the mean of all items on a scale excluding the person with the missing value being replaced. This approach retains item variability and has been shown to outperform other forms of substitution under various conditions (Raaijmakers, 1999). To calculate parameter estimates, the resulting option and test characteristic curves, and model comparison fit statistics, the default Bock-Aitkin (Bock & Aitkin, 1981) marginal maximum likelihood procedure of IRTPRO (Cai, Thissen, & duToit, 2011) was used. This estimation method is a full information maximum likelihood approach (Forero & Maydeu-Olivares, 2009; Wu & Bentler, 2013) and thus appropriately incorporates cases with missing data into each analysis and adequately accounts for MAR data (Enders & Bandalos, 2001).

Results

Tests of Assumptions

Prior to conducting item analyses, several preliminary analyses were conducted to ensure that IRT analyses were appropriate. First, the assumption of unidimensionality was examined by conducting Exploratory Factor Analyses (EFA) on each of the subscales of the ACES-TF. To appropriately analyze ordinal scale items, the default MPlus (Muthén & Muthén, 2012) robust weighted least square estimator (WLSMV) estimator was used. These EFAs were conducted by subscale and employed criteria from Reeve et al. (2007) to evaluate whether results supported
unidimensionality. Results were considered to support unidimensionality if the ratio of the first to the second eigenvalue exceeded 4. These criteria were met in all cases as ratios of the first to second eigenvalue ranged from 11.95 to 32.75. Scree plots also supported unidimensionality in all cases. Thus, the assumption of unidimensionality was retained for all subscales.

Also, the assumption of local independence was assessed through examination of the local dependence $\chi^2$ values (Chen & Thissen, 1997) generated by IRTPRO (Cai et al., 2011). According to these indices, there was evidence for local dependence on multiple scales of the ACES-TF. These instances of local dependence were addressed during item analysis to ensure that the resulting SMEIs did not evidence the same issues with local dependence.

Model Comparison

Next, different polytomous IRT models were tested to determine which model produced the best fit and should be used for item analysis and selection. Nested models were compared with $\chi^2$ difference tests and non-nested models were compared with comparative fit indices, such as the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978). In all cases the GPCM outperformed the PCM as indicated by statistically significant $\chi^2$ difference values. Furthermore, AIC and BIC values favored the GRM over the GPCM for all models. For these reasons, the GRM was used for all IRT analyses.

Item Analysis

Items were evaluated based on several criteria. First, the information curve of each item was evaluated when deciding whether to retain that item. Information levels at the $\theta$ range of -1.5 to -0.5 were emphasized given the projected importance of identifying those students in the “at-risk” range. Content validity was also considered when developing scales to ensure full content coverage. In addition to these two primary criteria, SMEIs were free of local
dependence and fit the employed IRT model at least adequately (as evidenced by RMSEA values lower than .10; MacCallum, Browne, & Sugawara, 1996; Table 2). All IRT analyses were conducted with IRTPRO version 2.1 (Cai et al., 2011). As a result of these procedures the ACES-TF was reduced by 56% from 73 to 32 items. Finally, due to initial content consideration, one item (Pays attention in class) was moved from the Study Skills subscale to the Engagement subscale. The full list of ACES-TF Items and items selected for the SMEIs can be found in Appendix A.

Table 2. GRM Model RMSEAs by ACES-TF and Set of Maximally Efficient Items (SMEIs) Subscales

<table>
<thead>
<tr>
<th>Subscales</th>
<th>ACES-TF</th>
<th>SMEI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic Skills</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading/Language Arts</td>
<td>.13</td>
<td>.09</td>
</tr>
<tr>
<td>Mathematics</td>
<td>.08</td>
<td>.07</td>
</tr>
<tr>
<td>Critical Thinking</td>
<td>.07</td>
<td>.07</td>
</tr>
<tr>
<td>Academic Enablers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpersonal Skills</td>
<td>.06</td>
<td>.04</td>
</tr>
<tr>
<td>Engagement</td>
<td>.05</td>
<td>.06</td>
</tr>
<tr>
<td>Motivation</td>
<td>.05</td>
<td>.07</td>
</tr>
<tr>
<td>Study Skills</td>
<td>.04</td>
<td>.03</td>
</tr>
</tbody>
</table>

Note. Reading/Language Arts RMSEA reported included only 10 items because the inclusion of all 11 items led to severe numerical errors.

Preliminary Reliability/Precision Analyses

Once the item analysis process was completed, preliminary reliability/precision analyses were conducted on the SMEIs of each subscale. Specifically, Test Information Functions (TIFs) were examined for each subscale and comparisons were made between ACES-TF and SMEI subscale TIFs. These TIFs indicate the total information resulting from the sum of item information values across θ. Information values in the SMEI subscales were necessarily lower than in the ACES-TF subscales. Nevertheless, all subscales retained the shape of their full form counterparts. Furthermore, using a formula that converts information to a standard reliability
metric ranging from 0 to 1 (reliability = 1-[1/information]; Petrillo, Cano, Mcleod, & Coon, 2015) there were very few instances in which the reliability estimate of the SMEI score dropped below .90, which has been identified as a lower-bound reliability threshold for individual decision making (Salvia, Ysseldyke, & Bolt, 2010), within the θ range of -2 to 2 for Academic Skills subscales and -2.5 to 1 for Academic Enablers subscales. Comparison TIFs for each subscale can be found in Figures 1 and 2.
Figure 1. Comparison of TIFs for ACES-TF and SMEI Academic Skills Subscales.

Note. TIF = Test Information Function; ACES-TF = Academic Competence Evaluation Scales – Teacher Form; SMEI = Set of Maximally Efficient Items; TIF for ACES-TF Critical Thinking subscale excludes items intended for a restricted age range (Items 29-33).
Figure 2. Comparison of TIFs for ACES-TF and SMEI Academic Enablers Subscales.

Note. TIF = Test Information Function; ACES-TF = Academic Competence Evaluation Scales – Teacher Form; SMEI = Set of Maximally Efficient Item; TIF for ACES-TF Engagement subscale includes item moved from Study Skills subscale (Item 70); TIF for ACES-TF Study Skills subscale excludes items intended for a restricted age range (Items 72 & 73).
Discussion

There are several important points for discussion resulting from these analyses. First, ACES-TF scores demonstrated a high degree of precision across θ. Examining general trends across ACES-TF subscales, however, indicates that score precision decreases markedly as θ approaches -3 or 2.5 for Academic Skills subscales and 1.5 to 2 for Academic Enabler subscales. This pattern indicates that scores are not as precise when considering individuals with extreme ability (high or low) in academic skills or above average ability in academic enablers.

Such findings are especially instructive because of the potential current use of the ACES-TF with students who are experiencing significant academic struggles (e.g., formal psychoeducational evaluation). It is precisely in these circumstances that the equal application of a single reliability index would be the most problematic. For example, the calculated Cronbach’s α coefficient for the ACES-TF Critical Thinking subscale with the current sample is .98. This single reliability index is roughly accurate for an examinee with an average score (at θ = 0, information = 22.84, and associated reliability estimate = .96). In contrast, for a student who is struggling it is an overestimate of reliability (at θ = -2.7, information = 3.59, and associated reliability = .72). Thus, the results of the current study indicate that scores on several subscales of the ACES-TF may be less precise when the measure is used with students of lower (and higher) skill. This is especially the case for the Academic Skills subscales and, to a lesser extent, the Engagement and Motivation subscales.

Overall, the SMEIs retained a relatively high degree of precision given the large number of items deleted from each subscale. Despite the overall positive SMEI information trends illustrated in Figures 1 and 2, it is important to note that precision was still reduced relative to the full-length version of the ACES-TF. For example, the standard error for the full form ACES-TF Mathematics scale (standardization sample) is .2 when θ equals 0. This standard error value
jumped to .28 through the SMEI development process. Despite the relatively large percentage increase in standard error (45%), the information value at this level of θ (information = 12.28) still yielded a high reliability estimate of .92. Thus, the item deletion procedures utilized to develop the SMEIs did not compromise score precision from an absolute perspective, but should be kept in mind when conducting IRT analyses in general. Similar projects conducted with other rating scales may not be as successful as the current project depending on the information functions of the original scales.

A brief measure of academic competence informed by the SMEIs identified in this study would lend itself well to secondary tier applications within a multi-tiered system of service delivery. For example, such a brief measure could be used as a second gate for students identified by a universal screener. This type of measure would be advantageous in this context because it would either confirm or deny the presence of major difficulty in the areas of concern and identify initial target areas for further examination or intervention (although these target areas would need to be confirmed via other data sources). This type of information could inform the development of intervention groups and could serve as a baseline measure for some sort of progress monitoring (e.g., in a pre-post type application). Such a measure would not be well-suited, however, to formal psychoeducational evaluation or applications at the tertiary level of service delivery. For these purposes, the published ACES-TF would be more appropriate as it provides higher reliability for scores from its subscales and includes additional items to inform the identification of specific skill deficits or problem areas for intervention. A brief measure of the length described in this study would also be well suited for educational research.

These results illustrate the promise of IRT approaches to improving the measurement efficiency of existing rating scales. This promise has been identified by several others (e.g.,
Edelen & Reeve, 2006), but polytomous IRT has not been widely used for this purpose with behavior rating scales in education. Given the wide variety of potential measurement applications present in the current educational and research landscape, IRT methodology should be used to conduct similar analyses on other commonly used measures. Such analyses could help identify ranges of ability in which currently used measures do not function well and areas of potential measurement efficiency to address in further test development. Resulting modifications to existing forms could result in measures that are better adapted for particular uses within school or research contexts.

There are several limitations to the current study. First, all results are preliminary and further evidence should be examined with an independent sample. Also, the RMSEA value of the ACES-TF Reading/Language Arts subscale was not able to be calculated. Whatever problem that led to this error was resolved in the item analysis process as the SMEI Reading/Language Arts subscale was associated with an acceptable RMSEA. This may have been due to the more multidimensional nature of the Reading/Language Arts scale, which includes aspects of reading, oral language, and written language. Specifically, the original ACES-TF scale includes five items measuring reading skill, five items measuring writing skills, and only one item measuring oral language skills. Conducting a model omitting this single oral language item eliminated the severe numerical error and resulted in an RMSEA of .13, indicating that, even though the scale was essentially unidimensional, its conceptual multidimensionality potentially affected estimation. Another important limitation to this study was the inability to account for the multilevel structure of the data due to the lack of a teacher identification variable. This limitation is especially important because students were not only nested within classrooms for this study, but also nested within rater. Thus, dependency in the dataset would be affected by
common factors in student classrooms and factors associated with being rated by the same individual, which could compound levels of dependence between observations. Overall, the results from this study should be viewed as preliminary and the SMEIs should not be conceived as standalone measures until further developed and fully validated.

Overall, the SMEIs identified in this study display the potential for polytomous IRT models to improve the measurement efficiency of teacher rating scale measures. Such developments could allow for broader consideration of various constructs primarily measured via rating scales in various settings (e.g., research and practice). A measure based on these SMEIs would also hold promise to specifically expand consideration of academic skills and academic enablers for educational researchers and practitioners. The methods demonstrated in this study should be used for similar scale development projects in the future to address the many ways in which rating scale measures are used in research and practice.
Chapter 3

Development and Evaluation of a Brief Measure of Academic Competence

One of the major weaknesses associated with low measurement efficiency is its effect on the utility of measures in applied settings and in educational research. The ACES-TF includes 73 items, which may limit the ability of educational researchers and practitioners to use it widely. Specifically, the length of the ACES-TF is likely too long to be feasibly used at the primary and secondary level of service delivery within a multi-tiered system (Brady, Evans, Berlin, Bunford, & Kern, 2012) or in its entirety within educational research. As such, a brief measure of academic competence (BMAC) has the potential to be useful to educational practitioners and researchers. Such a BMAC could function well within a multi-tiered system of educational service delivery. Potential applications at this level include progress monitoring and intervention evaluation.

A BMAC also could facilitate research on academic enablers. Since its development, there have been multiple studies using the ACES-TF to measure academic enablers, but many of these studies only included subscales or composites of the ACES-TF rather than the entire measure (e.g., Lee, Lee, & August, 2011; McCormick et al., 2013; Oakes et al., 2012; Wilson, Liu, Keith, Wilson, Kermer, & Zimbo, 2012). This is potentially problematic given the interrelations between academic enablers (e.g., DiPerna et al., 2002; 2005). Furthermore, as mentioned above Brady et al. (2012) indicated that administration of the ACES-TF for a large number of students (e.g., for research purposes) could lead to “excessive teacher burden” (p. 433). A BMAC informed by the SMEIs from Chapter 2 would allow researchers to include more subscales or the entire measure to broadly examine the impact and importance of academic enablers. Such a BMAC could also reduce participant burden, free up time for other measures,
and reduce the error associated with participant fatigue or boredom (Galesic & Bosnjak, 2009). Thus, the development of a BMAC could provide a valuable resource for educators and educational researchers interested in measuring student academic enablers.

Short forms of measures often are developed by the selective elimination of items on a well-validated long form (e.g., Rammstedt & John, 2007). Such an approach, however, has been extensively criticized. For example, Smith et al. (2000) noted two particular problematic assumptions: 1) that reliability and validity evidence for a long form automatically transfers to a short form and 2) because a short form is briefer than a full form, it requires a less extensive validation process. Other problems of short form development include a failure to demonstrate that a short form retains content coverage for each factor of a long form, a failure to ensure that a short form reproduces the factor structure of its parent form, and a failure to demonstrate reliability evidence from an independent sample. These shortcomings threaten the utility of developed short forms and call for more thorough validation processes in short form development. Brief measures developed according to procedures in Chapter 2 are no exception to this rule and must be independently validated prior to their use (Anthony et al., 2016).

Purpose and Hypotheses

**Purpose.** Given the considerations described above, the purpose of this study was to validate a BMAC informed by the SMEIs identified in Chapter 2 with an independent sample. To achieve this purpose, multiple forms of evidence were examined including evidence for content validity, reliability, structural validity, and construct validity.

**Hypothesis 1.** The structure of the BMAC is consistent with the structure of the published ACES-TF (DiPerna & Elliott, 2000).
Hypothesis 2. Scores from the Academic Skills and Academic Enablers scales and subscales of the BMAC are associated with acceptable precision in the at-risk range (i.e., reliability greater than .90 in the -1.5 to -0.5 \( \theta \) range)

Hypothesis 3. Academic skills and enablers as measured by the BMAC are positively related to directly measured academic achievement.

Hypothesis 4. Academic skills and enablers as measured by the BMAC are positively related to social skills.

Hypothesis 5. Academic skills and enablers as measured by the BMAC are negatively related to problem behaviors.

Method

Participants

The sample consisted of 301\(^2\) second through sixth grade students. Participants for these analyses were drawn from the follow-up year of a larger project evaluating the efficacy of the Social Skills Improvement System – Classwide Intervention Program (SSIS-CIP; Elliott & Gresham, 2007). For this study, students ranged in age from 6.67 to 12.33 years (median = 8.83 years). Demographic characteristics of these participants can be found in Table 3.

---

\(^2\) After the deletion of one outlier described below.
Table 3. *Demographic Characteristics of Participants from the Pilot Sample (N = 301)*

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Girls (n = 148)</th>
<th>Boys (n = 152)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>81</td>
<td>74</td>
</tr>
<tr>
<td>Minority</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>Hispanic</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Grade</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>Third</td>
<td>24</td>
<td>28</td>
</tr>
<tr>
<td>Fourth</td>
<td>23</td>
<td>22</td>
</tr>
<tr>
<td>Fifth</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Sixth</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>Educational Status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General Education</td>
<td>95</td>
<td>86</td>
</tr>
<tr>
<td>Special Education</td>
<td>5</td>
<td>14</td>
</tr>
</tbody>
</table>

*Note.* Percentages do not sum to 100 in some cases due to missing data and/or rounding.

**Measures**

**Brief Measure of Academic Competence.** A 32 item brief measure of academic competence (BMAC) was developed based on polytomous IRT analyses of the ACES-TF standardization sample. These analyses identified sets of maximally efficient items (SMEIs) for each ACES-TF subscale. As a result, the BMAC includes three Academic Skills subscales (Reading, Mathematics, and Critical Thinking) and four Academic Enablers subscales (Interpersonal Skills, Engagement, Motivation, and Study Skills). Overall scales were also included based on the composite of the Academic Skills (for the Academic Skills scale) and Academic Enablers (for the Academic Enablers scale) subscale items.

**Social Skills Improvement System – Teacher Rating Scales.** The Social Skills and Problem Behaviors scales and subscales of the Social Skills Improvement System – Teacher Rating Scale (Gresham & Elliott, 2008a; SSIS-TRS) were used in this study. As reported in the
test manual, there is evidence for the reliability and validity of scores from the SSIS-TRS (Gresham & Elliott, 2008b). With regard to reliability, reported Cronbach’s α for scores from the Social Skills scale was .97 and ranged from .83 to .92 (median = .90) for the Social Skills subscales. Cronbach’s α was .95 for the Problem Behaviors scale and ranged from .78 to .93 (median = .88) for the Problem behaviors subscales. Stability coefficients of 2 to 87 day intervals for the Social Skills scale and subscales ranged from .68 to .85 (median = .81) and from .76 to .86 (median = .84) for the Problem Behaviors scale and subscales. As evidence for validity, scores from the SSIS-TRS correlated as expected with scores from various measures (e.g., the Behavioral Assessment System for Children – Second Edition; Reynolds & Kamphaus, 2004). This evidence has been corroborated by several other studies (e.g., Gresham, Elliott, Vance, & Cook, 2011; Gresham, Elliott, Cook, Vance, & Kettler, 2010). Because these data were available for all students, comparisons between validity coefficients for the primary and intermediate (6th graders were considered intermediate students for the purposes of this project) grades were conducted.

**STAR Reading and Math.** The STAR Reading and Math (Renaissance Learning Inc., 2007) assessments were also used in this study. These measures are computer adaptive tests focusing on the first through twelfth grades. Internal consistency estimates ranging from .89 to .91 (median = .90) and stability coefficients ranging from .82 to .89 (median = .83) for STAR Reading scores. Similarly, reliability coefficients for STAR Mathematics ranged from .79 to .83 (median = .81) and from .73 to .79 (median = .74) for internal consistency and stability coefficients respectively (U.S. Department of Education, 2010). Meta-analytic studies also have yielded validity coefficients ranging from .71 to .73 (median = .72) for STAR Reading scores.

---

3 Reported internal consistency estimates were for the 5- to 12-year old age group that corresponded with the age of the participants in the current study.
and from .63 to .65 (median = .64) for STAR Mathematics scores when correlated to related standardized achievement tests (U.S. Department of Education, 2010). These data were only collected with a subsample of participants. STAR Reading data were available for 165 third, fifth, and sixth grade participants and STAR Math data were available for 162 third, fifth, and sixth grade participants. Because of this restriction of data, comparisons between primary and intermediate grades as described above for SSIS-TRS scores were not made for STAR scores.

**ACES-TF Engagement and Motivation Subscales.** The full length Engagement and Motivation subscales of the ACES-TF were collected as part of the larger study. Cronbach’s α coefficients for these subscales ranged from .94 to .95 for the Engagement scale and from .98 to .97 for the Motivation scale across subsamples (Kindergarten through the second grade and third grade through the fifth grade; DiPerna & Elliott, 2000). Furthermore, stability coefficients were .92 and .96 for the Engagement and Motivation scales as reported in the test manual (DiPerna & Elliott, 2000). In addition to evidence for reliability, there is evidence for the validity of scores from these subscales. First, these subscales have been included in various examinations of the internal structure of the ACES-TF (DiPerna & Elliott, 1999; 2000) as well as research examining the interrelationships between academic enablers (e.g., DiPerna et al., 2004). Finally, scores from these subscales have been shown to be related to related constructs as expected (e.g., DiPerna & Elliott, 2000).

**Procedures**

The data for this study were collected in conjunction with the final round of data collection from a broader multi-year project evaluating the efficacy of the Social Skills Improvement System – Classwide Intervention Program (SSIS-CIP; Elliott & Gresham, 2007). Data were collected at the end of the final year of the project from 63 classrooms across 7
different schools in the Northeastern United States. Teachers were asked to rate several of their students on the BMAC developed for this project. All participants were treated in accord with ethical principles and all procedures were approved by the institutional review board prior to data collection. Teachers received monetary compensation for every student on which they completed ratings.

Approximately 10 weeks prior to the collection of pilot data for the BMAC, teachers were asked to complete the SSIS-TRS and ACES-TF Engagement and Motivation subscales, and children completed the STAR Reading and Math assessments. Trained research staff monitored teachers’ completion of rating scales and students were administered academic measures on laptop computers. The interval between collection of the SSIS-TRS and ACES-TF full form subscale data and collection of the BMAC data ranged from 41 to 86 days (median = 71 days). The interval between collection of the STAR Reading and Mathematics data and the collection of the BMAC data ranged from 49 to 104 days (median = 79 days).

In addition to the measures described above, a brief survey on the content coverage of BMAC subscales was developed. This content validation survey was sent to 68 researchers and practitioners, 19 of whom (28%) responded to the survey. The full survey and characteristics of the respondents can be found in Appendix B.

**Data Analysis**

Several data analytic techniques were used to examine scores from the BMAC. First, responses to the content validation survey were examined. Next, precision of BMAC scores was examined. Specifically, graded response models with all parameters constrained to equal those estimated from the standardization sample were tested. If these models fit, the original model was considered to hold in the current sample. If these constrained models did not fit, follow up
unconstrained models were estimated and relevant precision metrics (e.g., TIFs) were compared with TIFs from the standardization sample. Next, a Confirmatory Factor Analysis (CFA) was conducted to evaluate whether the BMAC retained the structure of the original ACES-TF. To account for the nested nature of the data, results were adjusted for clustering within classrooms. Because item level data were ordinal, all analyses were conducted with polychoric matrices with the default robust weighted least square estimator (WLSMV) in MPlus 7 (Muthén & Muthén, 2012). The published structure of the ACES-TF was tested according to standard overall fit indices produced by MPlus including the Root Mean Squared Error of Approximation (RMSEA), the Comparative Fit Index (CFI), the Tucker Lewis Index (TLI), the $\chi^2$ value, and the Weighted Root Mean Square Residual (WRMR) value. To evaluate fit statistics, standard cutoffs recommended by Hu and Bentler (1999) were utilized. Specifically, RMSEA values near .06 and CFI/TLI values greater than .95 indicated adequate model fit. The WRMR statistic was examined, but considering its suboptimal performance (Muthén, 2008), other fit statistics were emphasized. Convergent validity analyses also were conducted with the measures described above, which consisted of computing correlations between scores from the BMAC scales and subscales and various scales and subscales of the measures described above.

**Results**

**Evidence for Content Validity**

Overall, expert reviewers generally rated BMAC subscales as sufficiently representing their intended content domain. The reviewers generally rated Academic Skills subscales lower than Academic Enablers subscales. Mean ratings for the Academic Skills subscales were 3.00, 2.68, and 2.63 for the Reading, Mathematics, and Critical Thinking subscales respectively$^4$.

---

$^4$ A rating of 2 indicated that the scale “somewhat” measured its intended construct and a rating of 3 indicated that a scale “sufficiently” measured its intended construct.
Mean ratings for the Academic Enablers scales were 2.95, 3.11, 3.05, and 2.79 for the Interpersonal Skills, Engagement, Motivation, and Study Skills subscales respectively. Despite varying mean ratings, for all but two subscales, at least 70% of expert raters considered the construct coverage “sufficient.” The two exceptions were the Mathematics subscale and the Critical Thinking subscale, for which 68% and 59% of raters considered the construct coverage “sufficient.” Some specific comments provided by respondents included questioning the exclusion of several items from subscales (e.g., computation from Mathematics, note taking from Study Skills) and noticing potential overlap among included items. Most of these comments involved items that had high levels of local dependence with other items. For example, the computation item from the Mathematics scale evidenced local dependence with most of the items on the Mathematics scale, such that including this item on the scale would have resulted in less than 3 items on the Mathematics subscale.

**Reliability/Precision**

To examine the reliability of the BMAC subscales, models in which all parameters were fixed to equal the parameters estimated from the standardization sample were tested. These models evidenced adequate fit (RMSEA < .10; MacCallum et al., 1996) for all but three subscales (Interpersonal Skills, Engagement, & Study Skills) for which unconstrained models then were tested. The Interpersonal Skills and Study Skills subscale unconstrained models evidence adequate fit (Figure 3); however, the unconstrained model for the Engagement subscale did not evidence adequate fit (RMSEA = .31).
Structural Validity

Data screening. Before conducting CFA analyses, data were screened and relevant assumptions were tested. Specifically, potential outliers were examined and the data were screened for normality. A case was deleted if its statistically significant Mahalanobis distance was associated with a leverage value that exceeded three times the average leverage value across items (Field, 2009). With regard to normality, items were considered nonnormal if they exhibited skew and kurtosis values greater than 2 and 7 respectively (Fabrigar, Wegener, MacCallum, &
Strahan, 1999). Finally, warnings potentially indicating excessive multicollinearity were also examined.

Results indicated that 24 cases were found to have statistically significant Mahalanobis values. Only one of these cases had a leverage value that exceeded 3 x the average leverage value, and this case was deleted for all analyses. No items were associated with excessive skew or kurtosis values. Generated warnings indicated excessive multicollinearity and cell sparsity related to several items. Further examination revealed sparsity in the cross tabulations of several items including 3 items on the Mathematics scale and 3 items on the Critical Thinking scale. To address these problems, categories were collapsed to ensure limited zero cells in these cross tabulations for the Mathematics scale. The problems with the Critical Thinking scale were able to be addressed by deleting one item that had sparse cells in its cross tabulation with the other two items. The resulting solution was not substantively different than the solution generated with all the items originally chosen for the BMAC and non-collapsed categories. Thus, the original model is reported to facilitate evaluation of the fit of the originally piloted measure, rather than a modified BMAC.

**Identification.** The structure tested in this analysis was a correlated factors design in which each subscale was represented by a factor, and all factors were allowed to intercorrelate. This approach was selected because prior structural analyses of similar scales, such as the ACES-TF, has been exploratory (e.g., DiPerna & Elliott, 2000) with oblique rotations consistent with a correlated factors design. This model was identified because it meets the two indicator rule (Kline, 2011) and factors were scaled automatically by MPlus.
Overall fit. The $\chi^2$ value of this model (Figure 4) was statistically significant; $\chi^2 = 1,060.41(443), p < .001$. The RMSEA associated with this model was .068 and the CFI and TLI values were both .99. Finally, the WRMR value was 1.17.

Loadings. Standardized loadings of items (Table 4) on their corresponding factors were high, ranging from .81 to 1.00 (median = .96). Interfactor correlations (Table 5) between Academic Skills factors ranged from .90 to .94 (median = .94). Interfactor correlations between Academic Enablers factors ranged from .71 to .86 (median = .77). Finally, interfactor correlations between Academic Skills and Academic Enablers factors ranged from .28 to .65 (median = .54).
Figure 4. Pilot Sample Correlated Factors Confirmatory Factor Analysis Model
<table>
<thead>
<tr>
<th>Item</th>
<th>Reading</th>
<th>Math</th>
<th>Critical Thinking</th>
<th>Interpersonal Skills</th>
<th>Engagement</th>
<th>Motivation</th>
<th>Study Skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.97</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.94</td>
<td></td>
</tr>
<tr>
<td>59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.94</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.95</td>
<td></td>
</tr>
<tr>
<td>61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.95</td>
<td></td>
</tr>
<tr>
<td>63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.84</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.95</td>
<td></td>
</tr>
<tr>
<td>67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.96</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.96</td>
<td></td>
</tr>
</tbody>
</table>
Table 5. *Interfactor Correlations from the Pilot Sample CFA (N = 301)*

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Reading/Language Arts</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Mathematics</td>
<td>.90</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Critical Thinking</td>
<td>.94</td>
<td>.94</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Interpersonal Skills</td>
<td>.33</td>
<td>.28</td>
<td>.35</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Engagement</td>
<td>.51</td>
<td>.50</td>
<td>.56</td>
<td>.71</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Motivation</td>
<td>.62</td>
<td>.58</td>
<td>.65</td>
<td>.79</td>
<td>.85</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>7. Study Skills</td>
<td>.63</td>
<td>.53</td>
<td>.62</td>
<td>.75</td>
<td>.76</td>
<td>.86</td>
<td>-</td>
</tr>
</tbody>
</table>

**Convergent and Discriminant Validity Analyses**

**Social skills and problem behaviors.** A general pattern emerged in which scores from BMAC Academic Skills scales and subscales were less strongly related to SSIS-TRS scores than scores from BMAC Academic Enablers scales and subscales. Across subsamples, SSIS-TRS Social Skills scales and subscales, and BMAC Academic Skills scales and subscales, correlations ranged from .08 to .43 (median = .24). The same correlations across BMAC Academic Enablers scales and subscales ranged from .30 to .80 (median = .59). A similar pattern was evident from examining SSIS-TRS Problem Behavior scale and subscale correlations. Across subsamples, SSIS-TRS Problem Behaviors scales and subscales, and BMAC Academic Skills scales and subscales, correlations ranged from -.38 to -.07 (median = -.24). The same correlations across BMAC Academic Enablers scales and subscales ranged from -.80 to -.27 (median = -.54).

Specific correlations are displayed in Tables 6 and 7.

Most relationships between BMAC scores and SSIS-TRS Social Skills and Problem Behavior scales and subscales were similar compared across grade level groups (primary and intermediate). The magnitude of these correlations was generally higher for intermediate level students. Furthermore, 16 correlations failed to reach statistical significance in the primary sample, but only 4 correlations failed to reach statistical significance in the intermediate sample.

Specific subscales of the SSIS-TRS that had several statistically non-significant correlations
included the Assertion, Empathy, Engagement, and Internalizing subscales, for which 3 of 4 BMAC Academic Skills correlations were statistically non-significant in the primary sample. The subscale with the most statistically non-significant correlations was the SSIS-TRS Bullying subscale, which had no statistically significant correlations with any BMAC Academic Skill scale or subscale in either sample. Despite these low statistically non-significant correlations with BMAC Academic Skills scale and subscale scores, all correlations between scores from these SSIS-TRS scales and subscales and BMAC Academic Enablers scales and subscales were statistically significant.

There were some statistically significant differences between samples with regard to the magnitude of correlations. The most consistent pattern emerged in the domains of the SSIS-TRS Engagement and Internalizing subscales. Specifically, SSIS-TRS Engagement scores were statistically significantly more strongly related to BMAC Total Academic Skills and Mathematics scores for intermediate elementary students. Also, SSIS-TRS Internalizing scores were statistically significantly more negatively related to BMAC Academic Skills, Reading, and Mathematics scores for the intermediate sample. The differences between the SSIS-TRS and BMAC Academic Skills scores were meaningful as well, as all correlations were small in the primary sample and moderate in the intermediate sample. There were several statistically significant differences involving BMAC Academic Enablers scale and subscale scores and SSIS-TRS subscale scale and subscale scores, but these differences were not as meaningful as the magnitude of these relationships was high for all correlations across samples.

**Academic achievement.** Scores from the BMAC Academic Skills scale were strongly positively correlated with both the STAR Reading and Math scores ($r = .56$ and .52 respectively). Also, scores from the BMAC Academic Enablers scale had a small to moderate
correlation with STAR Reading and Math scores ($r = .25$ and .33 respectively). Score correlations between BMAC Academic Skills subscales and STAR Reading ranged from .53 to .54 (median = .54). The same correlations ranged from .48 to .51 (median = .50) for STAR Math scores. For BMAC Academic Enablers subscales, these score correlations ranged from .12 to .33 (median = .20) for STAR Reading scores and from .21 to .37 (median = .29) for STAR Math scores. Although BMAC Reading scores were more highly correlated to STAR Reading ($r = .54$) than STAR Math ($r = .50$) scores, BMAC Mathematics scores were also more highly correlated with STAR Reading ($r = .53$) than STAR Math scores ($r = .48$).

**Engagement and motivation.** The score correlations between the BMAC Academic Skills scale was .50 for the ACES-TF Engagement subscale and .60 for the Motivation subscale. BMAC Academic Skills scale score correlations with the full form ACES-TF Engagement subscale ranged from .42 to .51 (median = .50). The same score correlations ranged from .53 to .58 (median = .56) for the full form ACES-TF Motivation subscale. The BMAC Academic Enablers subscale produced a correlation of .70 with the full form ACES-TF Engagement subscale and .80 with the ACES-TF full form Motivation subscale. Correlations between scores from the BMAC Academic Enablers subscales and scores from the full form ACES-TF Engagement subscale ranged from .45 to .76 (median = .61). Corresponding score correlations for the full form ACES-TF Motivation subscale ranged from .60 to .83 (median = .67). Scores from the BMAC Motivation subscale were more strongly correlated with scores from the full form ACES-TF Motivation subscale ($r = .83$) than the full form ACES-TF Engagement subscale ($r = .68$). This pattern also emerged for the BMAC Engagement subscale, which had scores that were more related to scores from the full form ACES-TF Engagement subscale ($r = .76$) than scores from the full form ACES-TF Motivation subscale ($r = .67$).
These relationships were largely similar across the primary and intermediate elementary grades. Score correlations across BMAC scales and subscales and the ACES-TF Engagement subscale ranged from .40 to .71 (median = .49) for primary elementary students and from .44 to .80 (median = .55) for intermediate elementary students. Similarly, score correlations across BMAC scales and subscales and the ACES-TF Motivation subscale ranged from .49 to .81 (median = .55) for primary elementary students and from .56 to .83 (median = .65) for intermediate elementary students. One correlation was statistically significantly greater for the intermediate sample than the primary sample (the correlation between the BMAC Engagement subscale and the full form ACES-TF subscale), but this correlation in both samples was high (.71 and .80 for the primary and intermediate samples respectively).
Table 6. *Correlations between BMAC Scale and Subscale Scores and SSIS-TRS Scale and Subscale Scores for Primary Elementary Students (N = 145)*

<table>
<thead>
<tr>
<th>SSIS-TRS Scale/Subscale</th>
<th>Academic Skills</th>
<th></th>
<th>Academic Enablers</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Reading</td>
<td>Math</td>
<td>Critical Thinking</td>
<td>Total</td>
<td>Interpersonal Skills</td>
<td>Engagement</td>
</tr>
<tr>
<td>Social Skills</td>
<td>.23</td>
<td>.23</td>
<td>.17</td>
<td>.26</td>
<td>.72</td>
<td>.77</td>
<td>.54</td>
</tr>
<tr>
<td>Communication</td>
<td>.24</td>
<td>.24</td>
<td>.19</td>
<td>.26</td>
<td>.66</td>
<td>.70</td>
<td>.54</td>
</tr>
<tr>
<td>Cooperation</td>
<td>.28</td>
<td>.30</td>
<td>.22</td>
<td>.29</td>
<td>.73</td>
<td>.74</td>
<td>.50</td>
</tr>
<tr>
<td>Assertion</td>
<td>.15</td>
<td>.15</td>
<td>.10</td>
<td>.18</td>
<td>.49</td>
<td>.40</td>
<td>.52</td>
</tr>
<tr>
<td>Responsibility</td>
<td>.24</td>
<td>.25</td>
<td>.18</td>
<td>.27</td>
<td>.70</td>
<td>.77</td>
<td>.45</td>
</tr>
<tr>
<td>Empathy</td>
<td>.14</td>
<td>.14</td>
<td>.08</td>
<td>.17</td>
<td>.60</td>
<td>.68</td>
<td>.44</td>
</tr>
<tr>
<td>Engagement</td>
<td>.16</td>
<td>.15</td>
<td>.12</td>
<td>.20</td>
<td>.61</td>
<td>.64</td>
<td>.54</td>
</tr>
<tr>
<td>Self-Control</td>
<td>.18</td>
<td>.17</td>
<td>.13</td>
<td>.21</td>
<td>.58</td>
<td>.73</td>
<td>.32</td>
</tr>
<tr>
<td>Problem Behaviors</td>
<td>-.22</td>
<td>-.21</td>
<td>-.19</td>
<td>-.25</td>
<td>-.67</td>
<td>-.73</td>
<td>-.46</td>
</tr>
<tr>
<td>Externalizing</td>
<td>-.19</td>
<td>-.18</td>
<td>-.16</td>
<td>-.23</td>
<td>-.58</td>
<td>-.71</td>
<td>-.31</td>
</tr>
<tr>
<td>Bullying</td>
<td>-.10</td>
<td>-.07</td>
<td>-.10</td>
<td>-.12</td>
<td>-.43</td>
<td>-.55</td>
<td>-.27</td>
</tr>
<tr>
<td>Hyperactivity/Impulsivity</td>
<td>-.29</td>
<td>-.30</td>
<td>-.23</td>
<td>-.30</td>
<td>-.71</td>
<td>-.69</td>
<td>-.48</td>
</tr>
<tr>
<td>Internalizing</td>
<td>-.15</td>
<td>-.14</td>
<td>-.13</td>
<td>-.18</td>
<td>-.56</td>
<td>-.52</td>
<td>-.53</td>
</tr>
</tbody>
</table>

*Note.* All correlations > |.16| are statistically significant (p < .05)
Table 7. Correlations between BMAC Scale and Subscale Scores and SSIS-TRS Scale and Subscale Scores for Intermediate Elementary Students (N = 154)

<table>
<thead>
<tr>
<th>SSIS-TRS Scale/Subscale</th>
<th>Academic Skills</th>
<th></th>
<th></th>
<th></th>
<th>Academic Enablers</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total</td>
<td>Reading</td>
<td>Math</td>
<td>Critical Thinking</td>
<td>Total</td>
<td>Interpersonal Skills</td>
<td>Engagement</td>
<td>Motivation</td>
<td>Study Skills</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Skills</td>
<td>.37</td>
<td>.34</td>
<td>.30</td>
<td>.39</td>
<td>.79</td>
<td>.79</td>
<td>.65</td>
<td>.71</td>
<td>.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communication</td>
<td>.34</td>
<td>.33</td>
<td>.27</td>
<td>.36</td>
<td>.70</td>
<td>.75</td>
<td>.57</td>
<td>.61</td>
<td>.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cooperation</td>
<td>.42</td>
<td>.41</td>
<td>.36</td>
<td>.43</td>
<td>.80</td>
<td>.79</td>
<td>.62</td>
<td>.74</td>
<td>.69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assertion</td>
<td>.27</td>
<td>.21</td>
<td>.25</td>
<td>.33</td>
<td>.45</td>
<td>.30</td>
<td>.52</td>
<td>.41</td>
<td>.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Responsibility</td>
<td>.32</td>
<td>.32</td>
<td>.25</td>
<td>.32</td>
<td>.75</td>
<td>.80</td>
<td>.56</td>
<td>.67</td>
<td>.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empathy</td>
<td>.23</td>
<td>.23</td>
<td>.16</td>
<td>.26</td>
<td>.69</td>
<td>.69</td>
<td>.55</td>
<td>.62</td>
<td>.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engagement</td>
<td>.37</td>
<td>.34</td>
<td>.34</td>
<td>.39</td>
<td>.74</td>
<td>.69</td>
<td>.64</td>
<td>.69</td>
<td>.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-Control</td>
<td>.25</td>
<td>.24</td>
<td>.20</td>
<td>.28</td>
<td>.64</td>
<td>.74</td>
<td>.44</td>
<td>.54</td>
<td>.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problem Behaviors</td>
<td>-.36</td>
<td>-.35</td>
<td>-.33</td>
<td>-.35</td>
<td>-.69</td>
<td>-.76</td>
<td>-.54</td>
<td>-.62</td>
<td>-.69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Externalizing</td>
<td>-.27</td>
<td>-.27</td>
<td>-.23</td>
<td>-.25</td>
<td>-.63</td>
<td>-.78</td>
<td>-.42</td>
<td>-.55</td>
<td>-.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bullying</td>
<td>-.15</td>
<td>-.15</td>
<td>-.15</td>
<td>-.14</td>
<td>-.43</td>
<td>-.59</td>
<td>-.28</td>
<td>-.37</td>
<td>-.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hyperactivity/Impulsivity</td>
<td>-.37</td>
<td>-.35</td>
<td>-.33</td>
<td>-.36</td>
<td>-.73</td>
<td>-.80</td>
<td>-.55</td>
<td>-.66</td>
<td>-.73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internalizing</td>
<td>-.40</td>
<td>-.37</td>
<td>-.38</td>
<td>-.38</td>
<td>-.55</td>
<td>-.47</td>
<td>-.54</td>
<td>-.51</td>
<td>-.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. All correlations statistically significant (p < .05) except correlations between SSIS-TRS Bullying and BMAC Academic Skills scales and subscales.
Discussion

Summary of Results and Evaluation of Hypotheses

Content validity evidence. Overall, BMAC subscales were generally rated as having sufficient coverage of the constructs they intended to measure. Despite these ratings, several subscales were identified as having potential problems in the area of content validity. Specifically, the Mathematics, Critical Thinking, and Study Skills subscales all had mean ratings below 3 (Sufficient) and the lowest percentages of experts rating the BMAC subscales as sufficient or optimal. Most of the difficulties appeared to stem from the exclusion of items that experts thought were necessary for sufficient representation of the construct. For example, one expert wrote, “It is challenging to put sufficiently if computations are not included in the [BMAC Mathematics subscale].” This item was excluded because its inclusion lead to extensive local dependence among several items in the SMEI analyses of the previous study. This finding calls attention to the difficulty of balancing statistical and theoretical concerns in scale development. Future research should examine the extent to which each method (statistical versus expert driven) scale development performs when considering the psychometric characteristics and utility of brief rating scale measures.

Reliability/Precision analyses. In general, analyses supported the precision of scores from the BMAC. For most subscales, SMEI parameters estimated with the standardization sample held in the BMAC estimated with the pilot sample. Scores from these subscales are associated with the same level of precision as found in the SMEI development process. For the Interpersonal Skills and Study Skills subscales, SMEI parameters estimated by the standardization sample did not hold with the BMAC when estimated with the pilot sample. The fact that previously estimated parameters did not produce adequate fit the current sample could
be indicative of item parameter drift (Goldstein, 1983). Nevertheless, unconstrained models fit adequately for these subscales in the current sample, indicating that these subscales still functioned adequately. Specifically, relative to the .90 reliability criterion (calculated with the formula from Petrillo et al., 2015) established in the previous study, both the Interpersonal Skills and the Study Skills subscales evidenced a drop in precision in the $\theta$ range of 0.5 to 1. The Study Skills subscale also evidenced a drop in precision in the $\theta$ range of -3 to -2. It is important to note that the target area for measurement (the “at risk” range) was measured with practically equivalent precision for both of these scales. A weakness of the pilot sample was that neither a constrained nor unconstrained model fit with the Engagement subscale. This lack of fit could be due to the small sample size relative to that typically employed for polytomous models. Further research should be conducted to improve the Engagement subscale. Formal methods of examining item parameter drift (e.g., Wells, Hambleton, Kirkpatrick & Meng, 2014) would be especially helpful. Overall, Hypothesis 2 was supported for every BMAC subscale except the Engagement subscale.

**Structural validity analyses.** Overall, the published structure of the ACES-TF fit the data adequately when tested with the BMAC, but there were problems with sparse cells indicating potential item redundancy on the Mathematics and Critical Thinking scales. Despite these weaknesses, the overall acceptable fit statistics and high item loadings and interfactor correlations indicate that BMAC largely retained the structure of the original ACES-TF. Thus, Hypothesis 1 was partially supported.

**Convergent and discriminant validity evidence.** Convergent validity relationships were generally as expected, although the magnitude of some relationships varied across constructs. In general, BMAC Academic Skills scale and subscale scores evidenced small to
moderate positive relationships with social skills and small to moderate negative relationships with problem behaviors. Nevertheless, there were some statistically nonsignificant correlations, especially those that included the SSIS-TRS Assertion, Empathy, Engagement and Internalizing subscales in the primary sample and the SSIS-TRS Bullying subscale in both samples. Also, BMAC Academic Skills scale and subscale scores showed moderate positive relationships with academic achievement and ACES-TF Engagement and Motivation full form subscale scores. BMAC Academic Enablers scale and subscale scores showed large positive relationships with social skills and ACES-TF full form Motivation and Engagement subscales. Furthermore, BMAC Academic Enablers scale and subscale scores showed large negative relationships with problem behaviors. Finally, BMAC Academic Enablers subscale scores evidenced generally small positive correlations with academic achievement. These relationships were in line with expectations regarding the direction and magnitude of relationships. Thus, Hypothesis 3 was supported and Hypotheses 4 and 5 were largely supported.

With regard to differences across grade levels, the direction of these relationships was the same across primary and intermediate grade levels, but the magnitude of these relationships was generally higher for intermediate students. This trend was particularly evident when considering the relationship between SSIS-TRS Engagement (e.g., initiation of social relationships) and Internalizing and teacher rated academic skills. This could indicate that as students progress through school, the relationship between social skills, problem behaviors, and academic achievement becomes more pronounced. This trend may be especially evident when considering student’s social engagement and internalizing difficulties. Future studies should examine potential age moderation of the relationship between these constructs and academic achievement.
Although the BMAC performed well in general, there are several indications of potential construct underrepresentation and redundancy. First, multicollinearity due to cell sparsity in this sample generated MPlus warnings, potentially due to item redundancy. This could indicate that the reduction in items made it easier for teachers to rate students equally across items. Also, as mentioned above, BMAC Academic Skills scale scores were related similarly to STAR Reading and Math scores. Although BMAC Reading scale score was more strongly associated with both STAR Reading and STAR Math scores than BMAC Mathematics scores, differences among validity coefficients for all Academic Skills scales and STAR measures were very slight (ranging from .48 to .54). Thus, the possibility of construct underrepresentation in the BMAC development process is important to consider.

From a practical perspective, this construct underrepresentation need not be thought of as an exclusively negative result of short form development. Depending on the context for measurement, educators may be willing to sacrifice the “edges” of conceptual construct space to focus on the “core” of the construct of interest and efficiently measure that core. This possibility is especially relevant for screening situations in which measurement is focused more on identifying students at risk for difficulties rather than providing detailed analysis of strengths and weaknesses. With reference to the Academic Skills scale construct underrepresentation discussed previously, screening would focus on broadly defined academic skills rather than the subconstructs and specific skills measured by the full form ACES-TF. In this case, however, predictive validity would be of paramount importance. Because predictive validity was not examined in this study, it should be addressed prior to utilizing this short form in practice.

There is another possibility that could explain the higher levels of multicollinearity found in the BMAC. There are several assumptions tested with IRT models. One is the assumption of
unidimensionality, which states that all items included in a scale measure the same unidimensional factor. This assumption, usually conducted in the context of large item banks, may pose problems when applied with relatively smaller sets of items. Specifically, because there are less items to define the latent construct, content representation is more likely to be excessively influenced by a few very similar items. As such, even though the assumption of unidimensionality may hold due to the interrelatedness between items, the construct of interest may be underrepresented by applying IRT techniques. This underrepresentation would then favor items that measure a restricted version of the original construct more completely, leading to a situation in which similar items are selected for a short form or SMEI. Such a situation would increase rather than decrease item redundancy despite the examination of local dependence statistics.

For example, the ACES-TF Mathematics scale includes eight items. Three of these items make explicit reference to problems or problem solving. Even though the assumption of unidimensionality held for this subscale, it is possible that the inclusion of three items with explicit reference to problem solving disproportionately influenced nature of the latent construct such that the Mathematics construct reflected a problem solving focus. Not surprisingly, the three items that explicitly mention problem solving had the highest discrimination parameters (strongly influencing their information curves). Fortunately, local dependence and content consideration resulted in the inclusion of only 2 of the 3 items in the final Mathematics SMEI scale; however, these items actually represent a higher percentage of the BMAC than the full form due to removal of other items. Beyond statistical indications of potential construct underrepresentation, one expert noticed this trend with the Critical Thinking subscale, stating, “I think the [Critical Thinking BMAC subscale] is quite narrow with two of the items appear[ing]
to overlap conceptually to a large degree (drawing conclusions, generalizing).” It is possible that this aspect of IRT analyses is largely responsible for the construct underrepresentation and extreme multicollinearity of the SMEI Math and Critical Thinking subscales. This dynamic of using IRT for short form development should be further examined in future research.

Another potential explanation for the high levels of multicollinearity and high item loadings is the rater effect (halo/horns effect; Thorndike, 1920). It is possible that item scores reflect teachers’ general impressions of a student (positive or negative) rather than specific ratings of individual behaviors. Rater effects have been researched extensively (e.g., Nisbett & Wilson, 1977) and shown to affect teacher judgements of student behavior (e.g., Ysseldyke & Foster, 1978; Allday, Duhon, Blackburn-Ellis, & Van Dycke, 2011). In some ways, rater effects are inherent in teacher rating scales, as all teacher rating scales reflect an individual teacher’s perspective on behavior, attitudes, or attributes of students. Rater effects become problematic however, if it is so great that scores vary systematically either based on independent teacher effects (e.g., a teacher who had a tendency to rate behavior as problematic regardless of who she was rating) or based on an interaction between a particular teacher and student (e.g., a teacher who has a particularly negative view of an individual student). In such cases, score utility could be significantly affected as scores would be indicative as much of variation between teachers as students. Such a possibility should be explored further for the BMAC from this study and the ACES-TF in general.

**Limitations and Directions for Future Research**

There are several important limitations to consider relative to this study. First, although somewhat diverse, the current sample was not representative of the current United States population of children (O’Hare, 2011). Also, the sample was smaller than those typically
employed to conduct polytomous IRT analyses. Future research should examine the performance of the BMAC developed in this project with a larger and more diverse sample. The lack of fit for the Engagement subscale is also an important limitation that should be addressed with further scale refinement and research. Limitations also existed with validity measures, especially the lack of data for ACES-TF full form subscales other than Engagement and Motivation to compare to the BMAC scores. Future research should directly address this limitation by gathering BMAC and ACES-TF data for each subscale to examine how closely BMAC subscales measure constructs tapped by the ACES-TF. Finally, the interval between collection of SMEI ACES-TF data and validity measures data was longer than is ideal for examining concurrent relationships.

There are many potential avenues for future research resulting from this study. First, the BMAC could be used in its current form for research purposes, however, it is likely that the BMAC would need further validation prior to being used in practice. Future research should also supplement the convergent and divergent validity evidence collected as part of this study. Another major area of future research should be predictive validity, which is especially important for uses of rating scale measures in screening contexts. If the measure would be used in screening contexts, Receiver Operating Characteristic (ROC) curve analysis and conditional probability analysis would be particularly useful for establishing and evaluating screening cut points.

**Conclusion**

Overall, there is evidence that the BMAC developed and piloted in this study was generally successful at producing reliable and valid scores while retaining a factor structure consistent with the model of the ACES-TF. Scores from the BMAC subscales yielded
reliability/precision estimates generally high enough to support individual decisions, especially for students at risk. These scores also evidenced expected relationships with convergent and discriminant validity measures and were rated as generally sufficient with regard to content coverage by 19 expert reviewers. As stated above, this measure could likely be used for research purposes immediately, but further validation would likely be required for it to be used in practice.
Chapter 4

Conclusion

Academic skills and enablers are important factors promoting students’ success in schools (DiPerna & Elliott, 1999; DiPerna, 2006). There are many measures that assess these constructs, including the ACES-TF (DiPerna & Elliott, 2000). One characteristic of psychoeducational assessment tools especially important in the current educational context is the efficiency with which constructs are measured. An examination of the measurement efficiency of the ACES-TF has not been conducted, and the length of the ACES-TF may be prohibitive for application in certain educational and research contexts. Thus, the purpose of this study was to examine and improve the measurement efficiency of academic competence via the ACES-TF. This purpose was addressed with two studies. First, through the use of polytomous IRT, the measurement efficiency of the ACES-TF was examined and sets of maximally efficient items (SMEIs) were identified for each ACES-TF subscale. This information was used to inform the second study, in which a brief measure of academic competence (BMAC) informed by the results of the first study was independently piloted and validated.

There are several key findings of these two studies. The process used to examine the measurement efficiency of the ACES-TF in the first study has been utilized in previous studies (e.g., Anthony et al., 2016), but the application of these methods with the ACES-TF standardization sample showcases the promise of polytomous IRT for measurement efficiency in general in addition to providing information regarding the measurement efficiency of the ACES-TF. Results indicated that the ACES-TF functioned well in general, but that there were various areas in which measurement efficiency could be improved. The second study of the dissertation extended the methodology used in the first study by piloting a BMAC based on the SMEIs...
identified in the first study. Results indicated that parameters estimated with the standardization sample generally held for the BMAC subscales and that the BMAC generally produced scores with evidence of reliability/precision. Furthermore, the content and construct validity of score from the BMAC were examined as well. Results generally supported the validity of the scores from the BMAC. Thus, in general, the results of this study bolster evidence supporting the use of polytomous IRT methodology to examine and improve measurement efficiency of rating scales. Furthermore, the BMAC developed in the second study holds potential as a brief measure to be used to assess students’ academic skills and enablers in applied and research contexts.

Together, these studies have several broad implications for improving the measurement efficiency and validity of teacher rating scales. First, the results of the IRT analyses demonstrate an inherent weakness of scales developed within a CTT framework. Specifically, the emphasis on global indicators of reliability can obscure information about score precision across ability level ($\theta$). Although the results of this study indicated that in general, reliability was sufficient across a broad range of ability for the ACES-TF, such findings may not hold for other ratings scales. Thus, practitioners and researchers likely make decisions about these measures based on unintentionally misleading overall reliability estimates. Such problematic interpretations may be especially likely given the pattern of information decreases as underlying ability approached extreme levels. Because the use of many rating scales may occur with children whose ability would be expected to fall within extreme ranges, this issue is even more potentially problematic.

This problem is well-documented (e.g., Embretson, 1996), but bears repeating given the dominance of CTT in scale development and reporting of reliability information. Short form measures developed from longer measures are particularly vulnerable to this limitation because
they are likely to have large standard errors at certain levels of \( \theta \) given their low item count. Existing full and short form measures should be examined with IRT methodology to inform users of the precision of scores under varying circumstances or across varying levels of the underlying trait being measured.

Given the great importance of measurement efficiency in the current educational landscape, results from this dissertation underscore the need for different methods of scale development. As mentioned above, the development of short forms from existing forms using IRT methodology, while promising, suffers from limitations of its own. Specifically, unless a construct is adequately represented by many items, it is potentially more likely to result in construct underrepresentation as IRT analyses are applied. This potential weakness highlights the importance of systematic content validation similar to the analyses conducted as part of this study. It also calls for reexamination of the appropriate use of essential unidimensionality standards for constructs typically measured by rating scales. Specific research should focus on determining the dimensionality of the constructs measured by the ACES-TF. If the assumption of unidimensionality is found to be generally unrealistic for ACES-TF constructs, even if technically met, future research should explore the use of multidimensional IRT models with academic competence constructs.

To maximize the benefits of IRT based analyses, it is likely that large banks of items would be needed. This could be problematic, as it is possible that certain non-cognitive variables such as academic enablers do not lend themselves well to the development of large item banks. For example, the ACES-TF Engagement subscale only included 8 items on which to conduct IRT analyses. This could be a fundamental limitation of using IRT in these contexts that should be further examined. Nevertheless, scale developers should strive to develop more and varied
items to assess constructs of interest. If calibrated, these items could be used either to develop static forms tailored to an estimated ability level, or to develop Computer Adaptive Tests (CATs). Such advances could lead to an increased use of rating scales in a wider variety of settings and for a broader range of purposes. Thus, the results of these studies could be viewed as a bridge between CTT methods of scale development and potential future methods of maximizing measurement efficiency in rating scale measures using IRT methodology.

Furthermore, by highlighting the particular advantages of polytomous IRT for examining and improving the measurement efficiency of rating scale measures, the current studies should lead to a broader examination of other potential advantages an IRT framework might lend to the development and use of rating scales. For example, IRT provides extensive procedures for identifying and addressing differential item functioning, scaling, and linking. As little research has examined utilizing these techniques for teacher behavior rating scales, future research should establish the feasibility of such approaches prior to their being applied on a broad scale.

Ultimately, these studies provide insight regarding maximizing measurement efficiency for rating scale measures of academic competence and other constructs. The BMAC developed in this project could be used for research purposes and holds promise to inform future ACES-TF scale development for applied purposes. Finally, the preceding discussion highlights the promise of IRT to improve the measurement efficiency of rating scale measures. Given the current educational context, such developments could greatly facilitate the full consideration of students’ academic skills and academic enablers among other variables commonly assessed with teacher rating scales.
References


doi:10.1097/01.mlr.0000250483.85507.04


Appendix A.
List of ACES-TF Items (Items Chosen for SMEI in Bold)

Reading/Language Arts

1. Reading Comprehension
2. Word-Attack
3. Vocabulary
4. Identifying a main idea
5. Reading Fluency
6. Spelling
7. Punctuation
8. Grammar
9. Written Communication
10. Oral Communication
11. Drawing Conclusions from Written Material

Mathematics

12. Computation
13. Pattern Analysis
14. Measurement
15. Understanding of Spatial Relationships
16. Mental Math
17. Using numbers to solve daily problems
18. Breaking down a complex problem
19. Problem-Solving

Critical Thinking

20. Synthesizing related information
21. Drawing conclusions from observations
22. Comparing similarities or differences among multiple objects or ideas
23. Classifying objects or ideas into categories
24. Generalizing from information or experiences
25. Identifying patterns from information
26. Deciding among alternate solutions
27. Investigating a problem or issue
28. Developing a solution to a problem
29. Identifying specific principles and their applications
30. Analyzing errors in information or processes
31. Constructing support for or against an issue
32. Analyzing supporting and opposing viewpoints
33. Testing hypotheses

Interpersonal Skills

34. Follows classroom rules
35. Corrects inappropriate behavior when asked
36. Expresses dissatisfaction appropriately
37. Accepts suggestions from teachers
38. Works effectively in a large group activity
39. Interacts appropriately with adults
40. Listens to what others have to say
41. Gets along with people who are different
42. Works effectively in a small group activity
43. Interacts appropriately with other students

Engagement

44. Speaks in class when called upon
45. Asks questions about tests or projects
46. Participates in class discussions
47. Volunteers answers to questions
48. Assumes leadership in group situations
49. Volunteers to read aloud
50. Initiates conversations appropriately
51. Asks questions when confused
70. Pays attention in class\(^5\)

Motivation

52. Is motivated to learn
53. Prefers challenging tasks
54. Produces high quality work
55. Critically evaluates own work
56. Attempts to improve on previous performance
57. Makes the most of learning experiences
58. Persists when task is difficult
59. Looks for ways to academically challenge self
60. Assumes responsibility for own learning
61. Is goal-oriented
62. Stays on task

Study Skills

63. Completes homework
64. Corrects own work
65. Finishes class work on time
66. Prepares for tests
67. Prepares for class
68. Turns in homework on time
69. Takes care of materials (e.g., textbooks, desk)
71. Completes assignments according to directions
72. Takes notes in class
73. Reviews materials

\(^5\) Item originally on ACES-TF Study Skills subscale.
Appendix B.

Content Validation Survey

Thank you for your willingness to provide feedback on this brief measure of academic competence. Please answer the following questions.

Reading Scale

**Set of Maximally Efficient Items (SMEI)**
- Reading Comprehension
- Vocabulary
- Reading fluency
- Spelling
- Grammar
- Written communication

**Excluded Items**
- Word-Attack
- Identifying a main idea
- Punctuation
- Oral communication
- Drawing conclusions from written material

Please respond to the following questions.

<table>
<thead>
<tr>
<th></th>
<th>Poorly</th>
<th>Somewhat</th>
<th>Sufficiently</th>
<th>Optimally</th>
</tr>
</thead>
<tbody>
<tr>
<td>How well does the SMEI represent the Reading skill domain?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

Comments or suggestions regarding the Reading SMEI:
Mathematics Scale

Set of Maximally Efficient Items (SMEI) | Excluded Items
--- | ---
Pattern Analysis | Computation
Mental math | Measurement
Using numbers to solve daily problems | Understanding of spatial relationships
Problem-solving | Breaking down a complex problem

Please respond to the following questions.

<table>
<thead>
<tr>
<th></th>
<th>Poorly</th>
<th>Somewhat</th>
<th>Sufficiently</th>
<th>Optimally</th>
</tr>
</thead>
<tbody>
<tr>
<td>How well does the SMEI represent the Mathematics skill domain?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

Comments or suggestions regarding the Mathematics SMEI:
Critical Thinking Scale

Set of Maximally Efficient Items (SMEI)
- Synthesizing related information
- Drawing conclusions from observations
- Generalizing from information/experiences
- Developing a solution to a problem

Excluded Items
- Comparing similarities/differences among multiple objects/ideas
- Classifying objects or ideas into categories
- Identifying patterns from information
- Deciding among alternative solutions
- Investigating a problem or issue
- Identifying specific principles and their applications
- Analyzing errors in information or processes
- Constructing support for or against an issue
- Analyzing supporting and opposing viewpoints
- Testing hypotheses

Please respond to the following questions.

<table>
<thead>
<tr>
<th>How well does the SMEI represent the domain of Critical Thinking skills?</th>
<th>Poorly</th>
<th>Somewhat</th>
<th>Sufficiently</th>
<th>Optimally</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

Comments or suggestions to regarding the Critical Thinking SMEI:
Interpersonal Skills Scale

Set of Maximally Efficient Items (SMEI)
Corrects inappropriate behavior when asked
Works effectively in a large group activity
Interacts appropriately with adults
Listens to what others have to say
Interacts appropriately with other students

Excluded Items
Follows classroom rules
Expresses dissatisfaction appropriately
Accepts suggestions from teachers
Gets along with people who are different
Works effectively in a small group activity

Please respond to the following questions.

<table>
<thead>
<tr>
<th>How well does the SMEI represent the domain of Interpersonal Skills?</th>
<th>Poorly</th>
<th>Somewhat</th>
<th>Sufficiently</th>
<th>Optimally</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

Comments or suggestions regarding the Interpersonal Skills SMEI:
## Academic Engagement Scale

**Set of Maximally Efficient Items (SMEI)**
- Participates in class discussions
- Volunteers answers to questions
- Asks questions when confused
- Pays attention in class

**Excluded Items**
- Speaks in class when called upon
- Asks questions about tests or projects
- Assumes leadership in group situations
- Volunteers to read aloud
- Initiates conversations appropriately

Please respond to the following questions.

<table>
<thead>
<tr>
<th>How well does the SMEI represent the domain of Academic Engagement?</th>
<th>Poorly</th>
<th>Somewhat</th>
<th>Sufficiently</th>
<th>Optimally</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

Comments or suggestions regarding the Academic Engagement SMEI:
Academic Motivation Scale

**Set of Maximally Efficient Items (SMEI)**
- Makes the most of learning experiences
- Persists when task is difficult
- Looks for ways to academically challenge self
- Assumes responsibility for own learning
- Is goal-oriented

**Excluded Items**
- Is motivated to learn
- Prefers challenging tasks
- Produces high quality work
- Critically evaluates own work
- Attempts to improve on previous performance
- Stays on task

Please respond to the following questions.

<table>
<thead>
<tr>
<th>How well does the SMEI represent the domain of Academic Motivation?</th>
<th>Poorly</th>
<th>Somewhat</th>
<th>Sufficiently</th>
<th>Optimally</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

Comments or suggestions regarding the Academic Motivation SMEI:
Study Skills Scale

Set of Maximally Efficient Items (SMEI)
Completes homework
Finishes class work on time
Prepares for class
Completes assignments according to directions

Excluded Items
Corrects own work
Prepares for tests
Turns in homework on time
Takes care of materials (e.g., textbooks, desk)
Pays attention in class
Takes notes in class
Reviews materials

Please respond to the following questions.

<table>
<thead>
<tr>
<th>How well does the SMEI represent the domain of Study Skills?</th>
<th>Poorly</th>
<th>Somewhat</th>
<th>Sufficiently</th>
<th>Optimally</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

Comments or suggestions regarding the Study Skills SMEI:
Please answer the following demographic questions:

What is your gender?
○ Female
○ Male
○ Other (specify) ____________________

What is your ethnicity?
○ Hispanic or Latino
○ Not Hispanic or Latino

What is your race? (please mark all that apply)
○ American Indian or Alaska Native
○ Asian
○ Black or African American
○ Native Hawaiian or Other Pacific Islander
○ White
○ Other (specify) ____________________

What is your primary language?
○ English
○ Spanish
○ Other (specify) ____________________

Please indicate the highest level of education completed.
○ Master's/Specialist Degree
○ Doctoral Degree
○ Other (specify) ____________________

What is your educational discipline?
○ Educational Psychology
○ School Psychology
○ Special Education
○ Other (specify) ____________________

What is your current employment setting?
○ College/University
○ Independent Research Agency
○ Public School or Schools (K-12)
○ State Department of Education
○ Other (specify) ____________________
What is your primary role in your current employment setting?
- Administrator
- Faculty/Researcher
- Practitioner/Clinician
- Other (specify) ____________________

For how many years have you been working in your field?

What is your current certification/licensure status (please mark all that apply)
- Certified by State Department of Education
- Nationally Certified School Psychologist
- Licensed by State Board of Psychology
- Other (specify) ____________________

Please answer the following questions regarding your usage of the ACES:

How often have you use the ACES in your research/practice?
- Never
- Rarely
- Sometimes
- Often

In what setting(s) have you used the ACES? (check all that apply)
- Clinical setting
- Research setting
- K-12 School setting
- Other (specify) ____________________

In what grade level(s) have you used the ACES? (check all that apply)
- Primary Elementary (K-2)
- Intermediate Elementary (3-5)
- Middle School (6-8)
- High School (9-12)
- Postsecondary (12+)

With which of the following populations have you used the ACES? (check all that apply)
- General Education populations
- Special Education populations
- At-risk populations (e.g., students being evaluated for Special Education services)

With which of the following populations have you used the ACES? (check all that apply)
- Low SES populations
- Middle SES populations
- High SES populations
With which of the following populations have you used the ACES? (check all that apply)
☑ Rural populations
☑ Suburban populations
☑ Urban populations

Have you used the ACES with racially diverse populations?
☑ Yes
☑ No

Have you used the ACES with linguistically diverse populations?
☑ Yes
☑ No

If you have any other suggestions feedback regarding the ACES SMEIs (or the original measure), please provide them below.

Thank you for completing this survey. Your responses will be used to evaluate and further improve the ACES SMEIs developed in this dissertation. If you would like a synopsis of these results, please send an email to Chris Anthony (cja5171@psu.edu).
Respondents

To better evaluate the content validity of the BMAC, 68 individuals were contacted, of whom 28% \((n = 19)\) responded. Fifteen respondents were faculty/researchers, one was an administrator, and one was a practitioner. All respondents had earned their doctoral degrees. Ten respondents were female and all respondents were non-Hispanic with regard to ethnicity, white with regard to race, and all spoke English as their primary language. Fifteen researchers worked in the field of School Psychology, three worked in the field of Special Education, and one worked in the field of Educational Psychology. Seventeen respondents worked in a college or university setting, one worked in a Public School, and one worked in an Academic Medicine setting. Respondents had worked from 1 to 37 (mean = 15) years in their respective fields. With regard to ACES-TF usage, on a Likert scale ranging from 1 \((Never)\) to 4 \((Often)\), two respondents indicated that they never used the ACES-TF, four respondents indicated that they rarely used the ACES-TF, nine respondents indicated that they sometimes used the ACES-TF and four respondents indicated that they often used the ACES-TF. Respondents had used the ACES-TF with children in the Primary Elementary to the Postsecondary level. Respondents also indicated that they had used the ACES-TF with General Education, Special Education, and At-Risk (e.g., students being evaluated for Special Education services) populations, with Low, Middle, and High Socioeconomic Status populations, and with Rural, Suburban, and Urban populations. Finally, most respondents indicated that they had used the ACES-TF with racially and linguistically diverse populations.
Vita
Christopher J. Anthony

CURRICULUM VITAE

EDUCATION

Ph.D.  Candidate – School Psychology
The Pennsylvania State University (Degree anticipated Spring 2016)

M.Ed. – School Psychology
The Pennsylvania State University (APA accredited & NASP approved)

B.A. – Summa Cum Laude – Phi Beta Kappa honors – Psychology and Theology
University of Notre Dame

PUBLICATIONS AND PRESENTATIONS

Journal Articles/Book Chapters


Posters and Presentations


