The Development of a Social and Emotional Well-Being Scale Using ESEM and CFA: Synergistic Stories in Complex Models

Christopher Hughes Busath
Brigham Young University
The Development of a Social and Emotional Well-Being Scale

Using ESEM and CFA: Synergistic Stories

in Complex Models

Christopher Hughes Busath

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Richard R Sudweeks, Chair
Paul Caldarella
Cade T. Charlton
Joseph Olson

Educational Inquiry, Measurement, and Evaluation
Brigham Young University

Copyright © 2021 Christopher Hughes Busath
All Rights Reserved
School districts face unique challenges as they implement social and emotional learning (SEL) initiatives, particularly when choosing or developing a high-quality scale to measure non-academic competencies. Like collaborations with the CORE school districts described by West, Buckley, et al. (2018) and the Washoe County School District described by Davidson et al. (2018), Alpine School District (ASD) partnered with Brigham Young University (BYU) to develop a scale (80 items) that reflected their Vision for Learning framework. In this pilot study, I describe the collaborative and iterative process used to develop a shortened version of the ASD Social and Emotional Well-Being Scale Beta Form A (23 items), which was administered to 461 secondary level students in the Spring of 2021. I implemented a relatively novel approach of comparing the results from exploratory structural equation modeling (ESEM) with target rotation with the results obtained from the more traditional confirmatory factor analysis (CFA) as a part of the iterative process. The scores of the resulting shortened version achieved acceptable fit (CFI = .97, TLI = .96, SRMR = .03, RMSEA = .06), high factor loadings ($M = .80, SD = .09$), high reliability indices by sub-scale ($M = .94, SD = .03$), and measurement invariance across gender and school level. I discuss insights that resulted from this novel approach in the development process, and make recommendations for its use, specifically in the field of SEL measurement. I end by encouraging the collaborative efforts between practitioners and researchers as a way of increasing capacities within districts, facilitating larger scale research, and ensuring the usefulness of findings.

Keywords: structural equation models, factor analysis, measures (individuals), interpersonal competence, school culture, school surveys
ACKNOWLEDGMENTS

I would like to thank my family and those closest to me, for their continued love and support. Particularly, I would like to thank Julie, my wife, who has been kind enough to always listen to me. My kids have been patient with all the extra hours I have spent on this project, and I’m grateful for their prayers. My parents’ encouragement to take this big step in my life has been fundamental. I also want to acknowledge God’s hand in helping an English teacher learn statistics. He has been more than gracious.

Additionally, I would like to thank Dr. Richard Sudweeks, who has been a pillar in my education, and a mentor on this research journey. I’m grateful for his high expectations of me as a student, researcher, and scholar, but also his great example of intellectual humility. I would also like to thank the members of my committee for their patience and support as I worked through repeated drafts of this dissertation.

Finally, I owe so much to Vallen Thomas, Mellissa Bostwick, Kim Jones and other employees at Alpine School District who allowed me to learn with them on this project. Their time, expertise, and professionalism were invaluable. And though they may never read this, as a former teacher, I’m grateful for all those teachers who allowed the rough but necessary initial stages of our research to interrupt their classroom.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE PAGE</td>
<td>i</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>CHAPTER 1: Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Collaborating to Address Practical Challenges</td>
<td>1</td>
</tr>
<tr>
<td>Complex and Closely Related Constructs</td>
<td>3</td>
</tr>
<tr>
<td>Statement of Purpose</td>
<td>5</td>
</tr>
<tr>
<td>Research Questions</td>
<td>5</td>
</tr>
<tr>
<td>CHAPTER 2: Review of Literature</td>
<td>7</td>
</tr>
<tr>
<td>Social and Emotional Learning</td>
<td>7</td>
</tr>
<tr>
<td>Resilience</td>
<td>9</td>
</tr>
<tr>
<td>Bounce-Back</td>
<td>9</td>
</tr>
<tr>
<td>Perseverance</td>
<td>10</td>
</tr>
<tr>
<td>Self-Mastery</td>
<td>11</td>
</tr>
<tr>
<td>Confidence</td>
<td>13</td>
</tr>
<tr>
<td>Safety</td>
<td>15</td>
</tr>
<tr>
<td>Factor Analysis</td>
<td>17</td>
</tr>
<tr>
<td>Common Factor Model</td>
<td>17</td>
</tr>
<tr>
<td>Statistical Models</td>
<td>19</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1 Example of Factor Loading Table ................................................................. 31
Table 2 Global Fit Statistics for the Shortened SEWS Beta Form A ......................... 62
Table 3 CFA and ESEM Correlations Among Factors for the Shortened SEWS Beta Form A ............................................ 62
Table 4 ESEM and CFA Factor Loadings of the Shortened SEWS Beta Form A .............. 64
Table 5 Invariance for Gender and School Level CFA ................................................ 68
Table 6 Reliability Estimates by Subscale .................................................................. 68
LIST OF FIGURES

Figure 1 Comparison of EFA and ESEM Descriptions, Code, and Path Diagrams .................37

Figure 2 Model 1 ............................................................................................................................56

Figure 3 Path Diagram Depicting the Structure of the Shortened SEWS Beta Form A ...........66

Figure 4 Percent of Item Variance Explained by Factors ...............................................................77
CHAPTER 1

Introduction

Research on the improvement of public education continues to support social and emotional learning (SEL) initiatives as both an ethical imperative and a practical advantage (Gehlbach & Hough, 2018). Durlak et al. (2011) conducted a meta-analysis of over two-hundred studies that evaluated the impact of school-wide SEL interventions. They reported effect size estimates of improvement for (a) SEL skills (0.57), (b) attitudes (0.23), (c) positive social behavior (0.24), (d) academic skills (0.27) and a decrease in (a) conduct problems (0.22) and (b) emotional distress (0.24). After analyzing several intervention programs, Belfield et al. (2015) concluded that there was a two-dollar economic return for every dollar invested in SEL, as a conservative estimate. Research continues to support the inclusion and study of SEL outcomes in school settings.

Collaborating to Address Practical Challenges

Measuring SEL comes with unique challenges, some of which teachers and principals may be unprepared to handle. SEL constructs tend to be complex, ill-defined, and interrelated. Designing measures that disentangle and properly reflect these complexities can be difficult, and so higher quality measures tend to come from sources outside the classroom. However, the absence of practitioner input can result in measures that are too long, difficult to score, or require expertise to properly understand (McKown, 2019; Read et al., 2019). Additionally, if developed scales do not precisely relate to initiatives already in place, educators and administrators are forced to weigh the cost of either revising those initiatives to match the scale, or to develop a scale that matches their initiatives.
In response to these challenges, some districts have chosen to partner with other organizations such as research groups and universities. For example, the CORE Districts (or CORE), consist of eight California school districts that have partnered to improve student outcomes through shared learning. As part of a quality improvement effort, CORE takes account of student academic achievement and growth, student social and emotional competencies, and school/culture and climate (West, Buckley, et al., 2018). CORE worked with the (a) Collaborative for Academic, Social, and Emotional Learning (CASEL), (b) the John W. Gardner Center for Youth at Stanford, and (c) Transforming Education (TransformEd) to identify the initial SEL constructs of interest, as well as to curate, develop, evaluate, and revise scales to measure them (Meyer et al., 2018; West, Buckley, et al., 2018). This multi-year, large scale collaboration resulted in (a) more accurate SEL measures, (b) greater awareness of and capacity to apply principles of measurement among educators and policy makers (Gehlbach & Hough, 2018; West, Buckley, et al., 2018), (c) meaningful insights into the degree to which average trait levels changed over time (West, Pier, et al., 2018), and (d) increased evidence of relationships between SEL and academic outcomes (Claro & Loeb, 2019a; Claro & Loeb, 2019b).

The Washoe County School District (WCSD) in Nevada developed a social and emotional competence (SEC) scale in collaboration with CASEL and the University of Illinois-Chicago (Davidson et al., 2018). Like CORE’s collaboration, WCSD has continued to develop and revise their scale over a multiyear period using statistical models found in item response theory (IRT) and focus group interviews (Crowder et al., 2019). The researchers described this partnership as being mutually beneficial, allowing them to create a “cost-effective, feasible measure that aligned with local school needs while being guided by broader theory and literature about core [Social and Emotional Competencies]” (Crowder et al., 2019, p. 282). They also
noted that the remarkable “knowledge transfer that occurred between psychometric and substantive experts” lead to “innovations in the way the data was used and disseminated throughout the project” (Davidson et al., 2018, p. 104).

The results of these partnerships are arguably more valuable to both parties than either party could produce separately. This collaborative approach facilitated large scale, longitudinal data collection allowing researchers to answer elusive questions about trait stability and the generalizability across populations. The process also resulted in a multi-year, phased roll-out and iterative development, which increased practitioners’ capacities and buy-in at the district-level (West, Buckley, et al., 2018.) as well as among teachers and students (Davidson et al., 2018). Researchers from both projects highlighted the value of a continuous and collaborative approach to scale development and implementation.

In a similar spirit of cooperation, Alpine School District (ASD) in American Fork, Utah partnered with researchers from Brigham Young University in Provo, Utah to develop and evaluate a scale based on their district’s *Vision for Learning* framework, with the intent to revise it for further use.

**Complex and Closely Related Constructs**

One challenge in developing SEL scales is the closely related nature of conceptually distinct constructs, which can lead to models that do not adequately fit the data and constructs that lack discriminant validity. For example, a portion of ASD’s scale, the ASD Social and Emotional Well-being Scale Beta Form A (SEWS Beta Form A), consisted of items intended to measure the following constructs: (a) Safety, (b) Confidence, (c) Resilience, (d) Self-mastery, (e) Bounce-back, (f) Perseverance, (d) Self-awareness, and (h) Self-management. Bounce-back and Perseverance were presumed to be first-order factors of the second-order factor Resilience; Self-
awareness and Self-management were presumed to be first-order factors of the second-order factor Self-mastery.

During the scale development process, it was possible that traditional methods of factor analysis such as CFA would result in a lack of model fit due to inevitable overlap between constructs like Resilience and Confidence, despite the conceptual distinctiveness intended by the researchers. A lack of discriminant validity evidence would be exacerbated in scales where there are more constructs and fewer items, a quality likely to be reflected in school-level SEL scales, which are often broad rather than deep.

When investigating the dimensionality of complex and closely related constructs, some researchers have recently suggested comparing results from the more restrictive CFA to results from the less restrictive factor analysis approach called exploratory structural equation modeling (ESEM; Marsh et al., 2014). Because an ESEM is a more generalized CFA model, they can be considered nested and thus comparable (Marsh et al., 2014). This comparison gives researchers three insights: (a) the degree to which constraining cross-loadings in the CFA model produces inflated correlations among factors, (b) the degree to which some items, if any, load significantly onto non-target factors, causing local misfit, and (c) additional clarity regarding whether ESEM or CFA/SEM should be used in subsequent analyses, for example, to test measurement invariance of the scale (Asparouhov & Muthén, 2009). Together, these insights can inform researchers about problematic items, model re-specification options, and the overall robustness of the model.

The technique of comparing CFA and ESEM results has been applied in the assessment of scales used in the field of (a) personality research (Boffo et al., 2012; Marsh, Nagengast, & Morin, 2013; Neff et al., 2019), (b) child behavior research (Hukkelberg et al., 2018), (c) mental
health research (Joshanloo, 2016b, 2018) and (d) exercise science research (Garn & Webster, 2018; Hoffmann & Loughead, 2019). However, to my knowledge it has not been applied in the assessment or development of SEL scales, though it may be ideal for such a context. Exploring the value of such an approach may prove beneficial for the current study as well as instructive of its potential use to the wider community of scale development researchers.

**Statement of Purpose**

The purpose of this pilot study was threefold: (a) to investigate the psychometric properties of the SEWS Beta Form A in terms of model fit, reliability, and measurement invariance, (b) to reduce the total number of items based on the initial findings to create a shorter, more student-friendly version of the scale, and (c) to investigate the degree to which comparing results from the more parsimonious CFA approach with the less restrictive ESEM approach can inform researchers about the dimensionality of a scale.

**Research Questions**

This study was guided by the following research questions:

1. Based on the results of CFA and ESEM, how does the modeled factor structure of the SEWS Beta Form A compare with the a priori theorized factor structure?
   a. In what ways does the current version of the SEWS Beta Form A need to be revised?
   b. Which items, if any, can be deleted to produce a shorter version of the SEWS Beta Form A with an acceptable level of reliability?
   c. What evidence exists that a model-based estimate of reliability would be more appropriate than the Cronbach’s alpha coefficient for use in estimating the reliability of scores for the various factors.
2. To what extent do CFA and ESEM produce similar and/or dissimilar results in terms of the following:
   a. model-data fit statistics
   b. item loadings on factors
   c. correlations among factors
   d. modification indices

3. To what extent is there evidence of configural, metric, and scalar invariance?
CHAPTER 2

Review of Literature

Social and Emotional Learning

SEL curricula aim to enrich human relationships and elevate the quality of academic achievement. As Weissberg et al. (2015) summarized, “SEL programming involves implementing practices and policies that help children and adults acquire and apply the knowledge, skills, and attitudes that can enhance personal development, establish satisfying interpersonal relationships, and lead to effective and ethical work and productivity” (p. 6).

Additionally, social, and emotional competence has been associated with several desirable longitudinal outcomes, such as increased chances of post-secondary success, marital status, career success, and decreased risks of incarceration and drug use (Domitrovich et al., 2017).

Social and emotional competence (SEC) consists of an array of skills, attitudes, behaviors, beliefs, and knowledge. Currently, there are a variety of frameworks that compete to define which competencies are most valuable. For example, the Character Lab (2021) framework includes skills and attitudes such as gratitude, grit, growth mindset, purpose, self-control, and honesty. The Clover Model highlights just four SEL skills: active engagement, assertiveness, belonging, and reflection (Partnerships in Education and Resilience, 2021). In contrast, the American College Testing (ACT) Holistic Framework expands four constructs (core academic skills, cross-cutting capabilities, behavioral skills, and education and career navigation skills) into over fifty sub constructs (Camara et al., 2015). One of the more widely recognized SEL frameworks was developed by the CASEL. The CASEL (2021) framework emphasizes self-management, self-awareness, social-awareness, and relationship skills.
Some researchers have examined the overlap between frameworks. The Ecological Approaches to Social Emotional Learning (EASEL) Laboratory at the Harvard Graduate School of Education has developed a user-friendly taxonomy of forty different frameworks, hosted on the ExploreSEL (http://exploresel.gse.harvard.edu/) web platform, which dynamically relates domains, frameworks, and terms to each other. The creators established connections between constructs not by how they are named, but by the behaviors that define them. For example, the behavior of standing one’s ground under peer pressure falls under the construct of “negotiation/refusal skills” in the World Health Organization (WHO)’s Skills for Health framework, “self-efficacy” in the Basic Education Curriculum Framework (BECF)’s Core Competencies developed by the Republic of Kenya, “relationship skills” in the Building Blocks for Learning framework, and “resistance skills” in the Developmental Assets (12-18) framework (Ecological Approaches to Social Emotional Learning Laboratory, n.d.).

The proliferation of social and emotional skills and frameworks may be a result of different perspectives on what to call the skills, behaviors, and knowledge. This leads to what some refer to as the jingle fallacy, a single skill given two different names, or the jangle fallacy, two different skills that are given the same name (Kelley, 1927 as cited in Credé et al., 2017). However, it might also be an indicator of the interrelated nature of social and emotional skills, attitudes, and knowledge. As demonstrated by Explore SEL, when operationalized most frameworks, despite their unique wording, have substantial overlap in meaning.

For this reason, in the next sections of the literature review, I have named and defined the constructs as ASD described them, and then have attempted to adequately place them among similar constructs from the literature.
Resilience

Resilience, in the ASD framework, consists of two subconstructs: Bounce-back and Perseverance. Bounce-back refers to a student's ability to recover from typical but difficult events (e.g., “I feel OK about myself even if I get a bad grade,” and “I don’t let others’ unkind words make me feel sad for too long”). Perseverance refers to a student's ability to persist in situations that may be taxing on self-control (e.g., “I keep going, even when the schoolwork is not easy for me,” and “I stay motivated during the school day”). In essence, Resilience is defined as the degree to which students can endure the challenges of daily life, particularly in the context of school.

Bounce-Back

There are a variety of definitions of resilience in the literature. After synthesizing over 270 studies, Windle et al. (2011) defined resilience as the “process of negotiating, managing and adapting to significant sources of stress or trauma” (p. 2). Academic resilience has been defined as the “heightened likelihood of success in school and in other life accomplishments, despite environmental adversities, brought about by early traits, conditions, and experience” (Wang et al., 1993, p. 137). Martin and Marsh (2009) contrasted resilience with academic buoyancy, the former being the type of fortitude needed to recover from traumatic experiences and the latter being the general ability to deal with everyday challenges. Fong and Kim (2019) contend that academic buoyancy is a skill that is more widely applicable, whereas academic resilience would only be called on by those few who experienced acute, chronic, and severe challenges. The Bounce-back element of Resilience in the SEWS Beta Form A more closely resembles the construct of academic buoyancy than other definitions of resilience.
**Perseverance**

The second component of the SEWS Beta Form A Resilience scale is Perseverance. In education, perseverance has been most popularized by Duckworth et al. (2007) as a subcomponent of grit. Duckworth et al. pioneered the research on grit in order to investigate why some intellectually capable students struggled in school. Their findings suggested that Grit, a combination of perseverance and consistency of interest, accounted for four percent of the variance in success outcomes beyond IQ and the Big Five personality model’s construct of Conscientiousness. Among the critics of grit as a unique and meaningful construct, Credé et al. (2017) argues that the perseverance sub-construct of Grit predicts academic outcomes better than perseverance and consistency of interest together. Constructs similar to perseverance like self-control, self-regulation, effort regulation, and effortful control are sometimes associated with conscientiousness, one of the Big Five personality traits (Roberts et al., 2014). Duckworth et al. (2007) contends that what sets grit apart from other self-control types of measures is the longevity of it. Someone may have the self-control to “effectively control his or her temper, stick to his or her diet, and resist the urge to surf the Internet at work—yet switch careers annually” (p. 1089).

Another feature that differentiates these constructs from one another is specificity. Muenks et al. (2017) found that self-regulation items tended to be more domain specific (e.g., working hard in math class) rather than general (e.g., hard worker), and did a better job of predicting classroom grades. In sum, there is evidence that measures of students' abilities to persevere and exert self-control in the short term and in more domain specific environments are predictive of success in those environments.
The items in the Perseverance subscale of the SEWS Beta Form A attempt to capture endurance beyond a single class period, but not to the extent that Duckworth et al. (2007) implied in their definition of perseverance. For example, the SEWS Beta Form A item “I stick with my goals, even if they are hard” does not have any explicit indicator of a multi-year commitment, and yet it does indicate commitment beyond a single day. Similarly, “I stay motivated during the school day” requires more self-control than is typically required for a class period but is not the type of endurance required for multi-day commitments.

**Self-Mastery**

Closely related to Resilience in the SEWS Beta Form A is the construct of Self-mastery. ASD defines Self-mastery as a composite of Self-awareness and Self-management. Self-awareness is the degree to which students can identify their emotions (e.g., “I can name my feelings”), the awareness they have about shifts in their emotions (e.g., “I notice when my feelings change (happy to sad, angry to calm”), and awareness of how their emotions affect them (e.g., “I understand what my body does when I have feelings (happy, sad, scared, angry)”). Self-management is the degree to which students can regulate their emotions (e.g., “I control my voice and body while waiting for my turn (at the drinking fountain, getting called on, etc.)”). The combination of awareness and regulation of one’s emotions in a school setting is considered Self-mastery.

Self-awareness and Self-management are two components of emotional intelligence (EI) (Cherniss et al., 2006; Salovey & Mayer, 1990, p. 194). In their watershed article on emotional intelligence, Salovey and Mayer (1990) asserted that individuals who were emotionally intelligent were more flexible, creative, motivated, and ultimately, more capable of “[weaving] a warm fabric of interpersonal relations” (p. 194). There is evidence that emotional intelligence
has an impact in school and at work as well (Cherniss et al., 2006) However, these two constructs have been further developed into a four-model approach. In a meta-analysis consisting of 162 studies, MacCann et al. (2020) found that strategic components of EI had a bigger impact than perception components. For example, emotional understanding (knowledge about the words used to describe emotions, and the causes and results of emotions) and emotional management (regulating and responding strategically to emotions) had a larger association with academic achievement than emotional perception (ability to identify emotions in people's faces and voices) and facilitation (using emotions to inspire thinking). One caveat to this general finding is that emotional perception and facilitation seemed to have a larger impact on humanities outcomes, such as English and history courses, where understanding people matters.

This study also highlighted a distinction in how EI is measured—rating scale v. ability scale. Rating scales, consisting of self-report surveys or observations reports, did not predict standardized test scores but did predict grades, while ability scales, consisting of scenarios and simulations, do predict standardized test scores, but only marginally when IQ is considered. MacCann et al. (2020) interpreted these results to mean that when students report a high EI, their academic performance will be predictably better based on their ability to build relationships with teachers and peers. Thus, both the type of EI being measured and the way it is measured seem to matter.

The items in the ASD-SEW Beta A Scale reflect both emotional understanding (Self-awareness) and emotional management (Self-management). ASD has opted for a self-report or rating scale rather than ability scale.
Confidence

ASD has also defined an overall Confidence construct as the degree to which students believe in themselves and their abilities (e.g., “I believe I have strengths,” “I believe in myself,” “I believe I will succeed in life,” and “I am brave enough to try new things.”). Items were worded to reflect a general construct of confidence, rather than a domain specific confidence. For example, while students may vary in their confidence to perform well in math class, physical education, or art, ASD is interested in a general level of confidence.

There are several constructs that capture this type of overall confidence, all of which fall under the general idea of positive thinking. Related constructs cited in the literature include optimism, hope, locus of control, fate control, positive or negative attributional style, growth mindset, and grit (Anderson et al., 2016). One review of the literature suggested conceptual overlaps of dispositional optimism, general self-efficacy, and hope in terms of positive expectations, future outcome expectations, goal oriented thinking, as well as associations with physical and psychological well-being, positive emotions, life satisfaction, adaptive coping, and high levels of effort and achievement (Gavrilov-Jerković et al., 2014).

Optimism, hope, and general-self efficacy seem to be the most applicable cousin constructs to the items that constitute the Confidence construct in the SEWS Beta Form A sub-scale, considering their general forward-thinking nature. Optimism relates to the degree that one has a positive outlook on life (Scheier & Carver, 1992). In the ASD measure, this sense of optimism comes out in the item “I believe I will succeed in life.” In the literature, hope is differentiated from optimism by a sense of goal orientation. Active hope is defined as a combination of two unique perceptions: (a) the perception that one can begin and persevere in completing a goal, and (b) the perception that one can come up with meaningful strategies to
complete the goal (Snyder et al., 1996; Stajkovic, 2006). These are sometimes referred to as the “agency” and “pathways” components of hope. In the ASD measure, this sense of hope is reflected in the items “I am brave enough to try new things” and “I believe I can accomplish anything I set my mind to.” General-self efficacy refers to the overall confidence in one's own capacity, especially to cope with difficulty (Scholz et al., 2002). This self-confidence is reflected in the SEWS Beta Form A items “I believe I have strengths,” “I know what I’m good at,” “I believe in myself,” and “I am confident I can perform as well as other students in school.”

Significantly, new research has shown supporting evidence for a hierarchical model that better defines the relationships between optimism, hope, self-efficacy and resilience in the form of core confidence (Stajkovic, 2006; Stajkovic et al., 2015). I considered it a possibility that a smaller set of items that tap into these different elements would result in a similar construct that can be considered confidence. Core confidence manifests these qualities “in a person who knows what and how to do (agency and pathways of hope), believes that s/he can perform those tasks (efficacy), keeps positive outcome expectations (optimism), and feels that s/he can ‘bounce back’ if failure occurs (resilience)” (Stajkovic et al., 2015, p. 30). It should be noted that the original SEWS Beta Form A attempted to distinguish bounce-back type behavior from confidence while Stajkovic et al. (2015)’s definition subsumes it.

While the distinction between these positive thinking constructs can be nuanced, all of them seem to contribute to and be influenced by a variety of other student outcomes of interest. Core confidence, or the belief that one can accomplish the variety of tasks ahead is theorized to contribute significantly to self-regulation (Stajkovic et al., 2015). Additionally, core confidence predicts positive classroom performance ($\beta = .25, p < .01$) and life satisfaction ($\beta = .63, p < .01$). Ciarrochi et al. (2007) found that positive attribution style and hope were both correlated
positively to academic outcomes ($r = .27$ and $r = .24$ respectively) and joy ($r = .33$ and $r = .22$ respectively) and correlated negatively to hostility. Even in the presence of grade point average (GPA), ACT scores, positive thinking has a salient relationship with academic outcomes (Day et al., 2010; Gallagher et al., 2017; Snyder et al., 2002). In a meta-analysis of the impact of hope on academics, Marques et al. (2017) found that hope positively related to both GPA and non-GPA academics ($r = .22$ and $r = .14$). Hope had strong relationships with positive assets such as global self-worth ($\rho = .43$), positive affect ($\rho = .37$), coping ($\rho = .26$), optimism and life satisfaction ($\rho = .55$), and goal-directed thinking ($\rho = .38$), while having a negative relationship with depression ($\rho = -.39$) and other negative effects ($\rho = -.26$). In the moderator analysis portion of the study, school level was a statistically significant influence on the variability of correlations across studies, with elementary school students through high school students achieving higher effect size relationships between hope and academics and hope and positive assets than college level students. Measuring and improving student confidence could have a ripple effect that would benefit students both inside and outside academics, especially pre-college students.

**Safety**

ASD is also interested in learning to what degree students feel safe at school. This construct is considered a desired condition to be fostered rather than a skill to be developed in the ASD *Vision for Learning*. In the SEWS Beta Form A, Safety is measured along two dimensions: physical (e.g., “I have enough food and water”), and emotional (e.g., “Students bully me online (over the computer or on the phone).”)

Safety is a commonly used domain in the multidimensional construct of school climate (Bradshaw et al., 2014; Osher & Berg, 2017; Thapa et al., 2013). The U.S. Department of
Education (2013) defined safety as the degree to which there is a lack of violence, bullying, harassment or substance use. In a review of school climate models, safety was a synonym for orderliness and discipline (school rules, norms, patterns of interactions), and sometimes defined more directly as perceptions about physical, academic, and socio-emotional security (Haynes et al., 1997; Rudasill et al., 2018). In other words, some models measure safety as the lack of negative behaviors, some measure it as a product of school faculty behaviors and processes, and some measure it as a student perception. The items of the SEWS Beta Form A were written to measure safety based on student perception (e.g., “I feel safe at school”).

The U.S. Department of Education has identified safety as an important component of school culture and climate across party lines. U.S. Secretary of Education Arne Duncan justified $70 million worth of grants to support school climate transformations, saying that “If kids don’t feel safe, they can’t learn. It’s that simple” (U.S. Department of Education, 2014, September 23). The U.S. Department of Education (2013) concluded from their research that a positive school climate resulted in closing achievement gaps, greater teacher retention, and higher graduation rates. That same school year, the U.S. Departments of Education and Justice also provided schools with a Discipline Guidance Package, noting that equitable treatment was a key component of ensuring that students felt safe (U.S. Department of Education, 2014, January 8). General research trends indicate that improving school climate can result in a variety of gains, including higher student self-esteem, self-confidence, psychiatric health, and lower student substance abuse, absenteeism, sexual harassment, and student suspension (Haynes et al., 1997; Thapa et al., 2013). Haynes et al. (1997) also found that school climate might be more important for students of color and students from low socio-economic backgrounds.
Later, under the leadership of the opposing party, and in response to horrific events of school violence, the Federal Commission on School Safety (2018) was established. After their initial research, the Commission proposed a three-tier approach to safety (prevention, protection, and response). The Commission’s final report (2018) suggested that schools address character education and encourage a positive school climate as a preventative measure. This would include the promotion of peer-to-peer connections, teacher-to-student connections, and collaborative response to cyberbullying.

Students’ perception of safety at school has been a recurring theme. In the SEWS Beta Form A, items that attempt to measure safety address feelings of connection between students and adults as well as students and teachers. It also has an item that addresses cyberbullying. As a component of the larger multidimensional construct of school climate, student safety is worth measuring and improving.

**Factor Analysis**

**Common Factor Model**

Factor analysis is often used in scale development. In the common factor model introduced by Thurstone (1947), factors, or latent traits, are assumed to influence or explain the covariation among indicators, or manifest variables (Bandalos, 2018). Latent traits are theoretical entities that must be measured indirectly. For example, math ability cannot be measured with a ruler, but is instead estimated by collecting samples of performance (e.g., addition or subtraction problems on a math test). Latent traits are also assumed to differ between individuals. A battery of test or survey items that tap into a latent trait will result in scores that vary according to that individual’s ability (e.g., more able students will score higher, and less able students will score lower). Because items will differ from one another (e.g., difficulty, content, clarity, etc.), some
items may tap into the target latent trait to greater or lesser degrees than others, and some may tap into multiple traits in ways that are not obvious by examining the raw scores. However, factor analysis allows researchers to investigate “the number and nature of latent variables or factors” that account for the relationships among the raw scores (Brown, 2015, p. 10). It has been an invaluable tool in assessing the quality of instruments and assisting in the interpretation of scores.

Importantly, factor analysis assumes that latent traits can be related to raw scores in a linear fashion. The following formula is used to model an individual’s estimated ability level, or factor score, with their observed score in a form similar to that of a linear regression equation:

\[ Y_{pi} = \lambda_i \eta_p + \epsilon_i \]  

(1)

In this formula, raw scores for each person on each item \( (Y_{pi}) \), are a function of their estimated factor score or trait level \( (\eta_p) \), the degree to which the \( i \)th item \( (\lambda_i) \) successfully measures that trait in the form of a factor loading \( (\eta_p) \), plus the error \( (\epsilon_i) \) involved in the testing occasion and process. The error term in this equation is considered to be random, or unrelated variance, that is partitioned out from the common variance that makes up the factor. Consequently, the larger the factor loadings, the smaller the error term will be. The smaller the error term is, the more the observed score reflects an individual’s trait level. This is also true if additional meaningful terms are added. If additional factors are added to the model, more of the variability in the scores can be explained, and the error term may decrease. Once again, factor analysis is the investigation into the number and nature of factors that account for the common variance among raw scores.

Typically, the common factor model is broken down into two categories of factor analysis, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). In one sense,
both EFA and CFA represent types of statistical models and methods. In another sense, they represent unique approaches to factor analysis, with distinct practices, advantages, and limitations. The distinction in statistical model has traditionally aligned with the distinction in approach. EFA was primarily used for analysis that was hypothesis-generating, exploratory, data-driven, and inductive in nature. CFA was primarily used for analysis that was hypothesis-testing, confirmatory, theory-driven, and deductive in nature (Bandalos & Finney, 2019; Brown, 2015; Flora & Flake, 2017).

Exploratory Structural Equation Modeling (ESEM) has been described as a method that sits somewhere between these two statistical models, taking on some of the advantages of both (Marsh et al., 2014). However, while ESEM bridges this gap, the approaches to data analysis remains importantly distinct. For this reason, I introduce EFA and CFA in statistical terms, and then explicate how each model fits into exploratory or confirmatory approaches. This introduction will help highlight the strengths and limitations of using ESEM in a scale development context, and important considerations that researchers should be aware of when conducting factor analysis with this newer method.

**Statistical Models**

In purely statistical terms, EFA estimates the degree to which each item in the scale loads onto each factor in the model. In the estimation process, researchers’ influence on the specification of the model is limited. Researchers indicate how many factors are believed to account for the covariance among the items, whether those factors are correlated, what type of rotation to use, and the EFA software partitions the common variance automatically. While scale developers usually generate items to measure a particular construct, an EFA does not allow for such specification.
For example, consider a generic school test that contains sections of items to measure a student’s ability in mathematics and reading. The researcher conducting the EFA knows which items are intended to measure which latent trait but is limited to specifying that there are two factors that explain the covariance among the scores. Typically, with such distinct traits (e.g., a student can be competent at reading and relatively incompetent at math), scores obtained from math items will be more highly related to scores from other math items, and less related to scores obtained from reading items. An EFA will partition this common variance among the math scores into high loadings on one factor and low but non-zero loadings on a second factor. Similarly, items meant to measure reading in this EFA will load highly on the second factor but will still achieve non-zero loadings on the first factor. Because the user only specifies the number of factors, interpreting what the factors are will depend on item-factor loadings (reading items may load highly on Factor 1, and math items may load highly on Factor 2, or vice versa).

Another feature of EFA is the indeterminate nature of the factor loadings. EFA will initially produce a solution that maximizes the fit of the model to the data. However, the factor loadings themselves can be adjusted or rotated without changing the fit of the model (Bandalos, 2018). Rotational methods have been developed to maximize the interpretability of the factor loading matrix by approaching what Thurstone (1947) described as simple structure. Rotational methods that allow factors to be correlated (oblique) and methods that force them to be uncorrelated (orthogonal) are available (Bandalos, 2018, p. 331). The researcher’s choice of rotation will affect the interpretability of the factors and interpretation of the model.

In contrast to this unrestricted solution, CFA requires that the researcher specify the model beforehand. This specification usually identifies (a) which items are related to which factors, (b) whether factors are correlated with one another, and (c) whether errors terms are
correlated. In my example, items meant to measure reading would be allowed to load onto the specified reading factor, but their loadings on the math factor would be constrained to zero. This Independent Clusters Model CFA (ICM-CFA) was popularized by Jöreskog (1969), and software that models it has traditionally included other useful features. Common software used to conduct CFA such as Mplus and lavaan (Muthén & Muthén, 2017; Rosseel, 2012) allow for the testing of measurement invariance, reporting standard errors and model fit statistics, and allow for the integration of the CFA results into more complex structural equation models (Brown, 2015). These features are often highlighted as advantages of CFA over EFA (Bandalos, 2018; Brown, 2015). While there are ways to accomplish most of these tasks using EFA, they tend to be cumbersome (Marsh et al., 2014).

Both EFA and CFA are specific instantiations of the common factor model (Flora & Flake, 2017), with the more constrained model (CFA) being nested within the more general EFA model. This relationship allows researchers to statistically test the degree to which more parameters increase the fit of the more general model (Marsh et al., 2014). Another result of this relationship is that EFA models will often fit better than CFA models, because more factors are accounting for covariation in the item scores. In purely statistical terms, EFA and CFA take up opposing ends of the same spectrum: a priori input.

**Exploratory Approach**

While the least restrictive factor model is referred to as exploratory, and the most restrictive is referred to as confirmatory, reviews of how these statistical models are used reveal that both EFA and CFA are used for exploratory and confirmatory studies (Browne, 2001; Bandalos, 2018; Flora & Flake, 2017).
Often, EFA is more suited for the initial exploration and development of a scale. Flora and Flake (2017) suggest that starting with an EFA allows for the discovery of “unanticipated, but substantively meaningful” factors (p. 82). Because EFA results are heavily data-driven, they can reveal relationships between items and factors that are sometimes surprising, insightful, and diagnostically valuable. For example, when evaluating the internal structure of a survey with closely related constructs such as bounce-back and perseverance, an EFA may indicate that one item written to measure bounce-back instead loads highly on perseverance.

However, best practice associated with EFA and an exploratory approach involves a number of preliminary steps. First, the researcher must decide whether the nature of the data justifies a factor analysis by examining the correlation matrix, the Kaiser-Meyer-Olkin (KMO), and Bartlett’s test of sphericity (Bandalos, 2018; Howard, 2016). Then, if it appears that there is sufficient common variance to produce factors, researchers use the Kaiser criterion, skree plots, parallel analysis, or the Hull method to determine how many factors to model or retain (Bandalos, 2018; Howard, 2016; Lorenzo-Seva et al., 2011).

Once a range of factors to retain has been identified, using EFA to explore the scale provides unique insights. The input from the researcher amounts to identifying how many factors the model should estimate, and then data-driven methods called rotations are used on the resulting factor matrix to minimize the loadings of items on some factors and maximize loadings on others. Rotation methods attempt to cluster items on factors to assist in the interpretation of the factor structure. This approach is sensitive to shared variance between item scores that may have not been anticipated by the researchers. The rotated factor matrix could result in new understandings about constructs (e.g., that students respond differently to adults v. peers in terms of respect), or it may result in additional atheoretical nuisance factors that account for method or
distributional effects (e.g., items that are negatively worded v. positively worded, items that measure socially desirable traits, items that have similar distributions, etc.; Bandalos, 2018, p. 118). In a theory-generating process, researchers examine the factor loadings and attempt to determine what the factor represents by analyzing patterns within the content of the items.

In the exploratory process, researchers will need to determine the value of each model (1 factor model, 2 factor model, etc.), as well as the items within the models. Commonly, researchers in the literature examine the magnitude of factor loadings using a priori thresholds (.3, .32, .4, etc.), identify items that cluster around a factor, and choose models that are theoretically meaningful (Bandalos, 2018; Bandalos & Finney, 2019; Flora & Flake, 2017; Howard, 2016). A common framework is Thurstone’s (1947) simple structure, which gives empirical guidelines to evaluate the degree to which items cluster around a factor. The content of items that have secondary loadings above the a priori threshold would be scrutinized, and possibly revised or eliminated. Bandalos (2018) lamented the use of cross-loadings (higher than desired secondary loadings) to identify and eliminate items and suggests that researchers wait until cross-loadings are duplicated in an independent sample, claiming that factor structures can be unstable when loadings are small, or the sample is small. Without additional testing, the researcher will not know whether problematic loadings are the result of a poorly devised item or the result of idiosyncrasies within the sample. Additionally, the elimination of items that were theoretically valuable in measuring the construct may result in what the Standards for Psychological and Educational Testing describe as construct underrepresentation (American Educational Research Association et al., 2014, p. 10). However, when the researcher is determined to eliminate items, Bandalos and Finney (2019) recommend eliminating one item at a time (p. 106). They suggest that eliminating one poorly performing item could resolve
problematic issues for other items. After the item is eliminated, the factor analysis needs to be run again. Factor loadings and factor scores are not initially known, but are estimated from the covariance matrices, which in turn is derived from the observed scores of the items. When an item is removed, the correlation matrix changes, and the subsequent factor analysis results also change. In other words, the combination of items is as important as the items themselves.

At the same time, CFA is sometimes employed in exploratory ways. Despite the best efforts of researchers and scale developers, a model that should fit one sample based on theory or past empirical evidence may not fit. Unfortunately, because simple structure is implemented in the specification of the model, items that may be causing misfit are difficult to identify in CFA models because they are not allowed to cross-load. Instead, researchers use different methods to diagnose misfit. Researchers that are using CFA will consult a table of residuals or a modification index in a post-hoc modification process (Bandalos, 2018; Brown, 2015). The modification index estimates the degree to which adding or deleting a parameter will decrease the chi-square value, improving overall model fit. Some of the suggested alterations to the model may include switching an item to load on a different factor, or to correlate error terms between items. Correlating error terms to improve model fit without a strong justification is considered particularly “egregious,” because it indicates there are strong relationships between variables that are not being accounted for in the model (i.e., the model is incorrectly specified; Bandalos & Finney, 2019, p. 117). Instead of accounting for the relationship in a theoretically meaningful way, correlating error terms improves model fit without contributing to the overall interpretability of the scale. Once these methods are used to adjust or re-specify the model, what had been a hypothesis-testing analysis has become an exploratory or hypothesis-generating analysis (Bandalos, 2018).
MacCallum et al. (1992) warn specifically about the use of modification indices to improve model fit but do so in the context of data-driven model re-specification in general. They recognized that exploratory approaches are sometimes necessary and beneficial but insisted that all data-driven model specification efforts capitalize on chance (even changes motivated by cross-loadings from an EFA model). Specifically, MacCallum et al. warn that the degree of concern increases based on (a) the size of the sample, (b) the number of modifications, and (c) the interpretability of the modifications. MacCallum et al. cite MacCallum (1986) as demonstrating that longer model searches did not result in a higher rate of correct model selection. Browne (2001) also acknowledges that CFA can be used for exploratory purposes but suggests that the EFA model provides a more direct route to discovering where the model is misfitting in the form of rotated factor loadings. Browne also suggests that CFA is sometimes used in exploratory approaches because researchers erroneously think that CFA provides standard errors for factor loadings and EFA models with rotation do not. While thresholds for what passes as a meaningful factor loading are arbitrary and debatable (Howard, 2016), software such as Mplus has reported the standard error and statistical significance of loadings in EFA models since Version 5, published in November of 2007 (Mplus, n.d.). However, exploratory methods where models are repeatedly estimated from the same sample violates the underlying assumptions of statistical inference, inflating the Type I error rates (Bandalos & Finney, 2019). In other words, $p$ values obtained in data-driven re-specification cannot be interpreted as indicators of statistical significance, and inferences made about the generalizability of the corresponding parameter estimates to the population are tentative.
**Confirmatory Approach**

EFA and CFA are also both used for confirmatory approaches. In general, there are three qualitative differences between confirmatory and exploratory approaches: (a) the degree of a priori specification, (b) the use of statistical inference, and (c) the researcher’s response to initial findings (Bandalos & Finney, 2019; Flora & Flake, 2017). In an exploratory approach, researchers are justified and encouraged to adapt their hypothesis based on initial findings using the same sample (altering the model through addition, revision, or deletion), whereas confirmatory approaches are the test of how well the model fits without re-specification (Bandalos & Finney, 2019, p. 112). It is likely that multiple models could fit a data set, but to conserve the nature of the confirmatory approach, researchers would need to specify the models before-hand, and not as a response to the initial analysis (Bandalos & Finney, 2019).

Before CFA was widely accessible, personality theorists were using EFA to both explore and confirm their findings of the Big Five Model. McCrae et al. (1996) describes how researchers used natural language adjectives and a variety of personality questionnaires to develop the long-term validity argument for the five-factor personality model. The basic five factor structure was successfully fit to data collected across studies that sampled different populations (White and non-White, male and female) and across different measurement methods (self-report, observation, or from a spouse). While model fit statistics were not explicitly used in the studies that I was able to reference, these statistics are available for EFA models in Mplus and several R packages (e.g., lavaan, or psych; Muthén & Muthén, 2017; Revelle, 2017; Rosseel, 2012). As described before, standard errors for factor loadings and correlations between factors are also available in EFA models. When an EFA model is fit to a sample based on prior evidence
and theory, these statistics of global fit and local fit can serve as supporting empirical evidence of a hypothesis-testing, deductive, confirmatory analysis.

However, CFA has been the preferred model for confirmatory type approaches. Early in its development, CFA provided ways of statistically testing model fit (Bandalos, 2018). Additionally, the nature of running a CFA lends itself to a confirmatory process. Researchers must specify the model beforehand, identifying which items load onto which factors, which factors are correlated (if any) with other factors, what the scale will be, and if there are any correlated errors. To the degree that CFA requires more complex and specific a priori modeling, a well-fitting CFA model can be seen as more trustworthy. For example, a researcher using CFA could make the following claims through model specification:

1. All of $x$ items will load on factor $x$, but will not load on factor $y$, nor will they load on factor $z$.
2. All of $y$ items will load on factor $y$, but will not load on factor $x$, nor will they load on factor $z$.
3. All of $z$ items will load on factor $z$, but will not load on factor $x$, nor will they load on factor $y$.
4. Factor $z$ and $y$ will be correlated.
5. Factor $x$ will not correlate with factor $z$ or factor $y$.

In contrast, the hypothesis claims made when using EFA may amount to as little as the following:

1. All items will load onto $x$ number of factors.
2. All $x$ factors will be correlated.
The level of specification used in CFA models allow for stronger hypothesis-testing, where a complex prediction is proposed and then tested against empirical data. Additionally, even more complex analysis involving measurement invariance, longitudinal studies, or multiple indicators and multiple causes (MIMIC) models are other examples of how the nature of specifying the model lends itself to the confirmatory process (Bandalos, 2018; Brown, 2015). For example, if intelligence is suspected to influence grades and standardized tests in the presence of other variables like socio-economic status, a CFA of an intelligence scale can produce the latent scores for student intelligence free of error, which is then included as a predictor in the structural portion of the model. If the results match the theory, researchers have garnered more construct validity evidence that their scale is a measure of intelligence (i.e., predicts academic outcomes in the presence of other independent variables; American Educational Research Association et al., 2014, pp. 26-27).

Both EFA and CFA have and can be used in hypothesis-testing, deductive, and confirmatory studies. However, CFA has been rightfully preferred based on the level of model specification (i.e., theory driven), the statistical tests available, and the ease with which CFA models could be related to additional variables.

Exploratory Factor Analysis Rotation

The practice and methods of rotating factor loadings becomes another salient feature of ESEM and deserves further explanation. One complication with EFA is that the loadings are indeterminate, meaning that there are an almost infinite number of estimations for each loading that will return the same model fit statistics. In other words, the initial EFA solution is not the only solution. However, the relationship between items and factors is more easily interpreted if a cluster of items are primarily related to a single factor rather than multiple factors. Thurstone
(1947) outlined rules that would lead to more interpretable solutions. I have added examples in parenthesis for additional clarity. In a factor loading matrix with \( i \) number of rows (items) and \( f \) number of columns (factors):

1. Each row contains at least one zero (every item should fail to load on at least one of the factors).
2. Each column should contain at least \( f \) number of zeros (if there are five factors, then there should be at least five zeros in each column).
3. Every pair of columns should have several rows with a zero in one column but not in the other.
4. If \( f \geq 4 \), every pair of columns should have several rows with zeros in both columns.
5. Every pair of columns of the matrix should have few rows with non-zero loadings in both columns.

If the factor matrix can be rotated to accommodate the first rule, then there will be greater simplicity and interpretability; if it can be rotated to accommodate the other rules then there will be greater stability (Yates, 1997, as cited in Browne, 2001).

The factor loadings in a two-factor model can be visualized on a coordinate plain in which the y-axis and the x-axis are factors, and the distance from a variable to each axis relates to the corresponding factor loading. If the axes are rotated, the distance between the variables and the axes will diminish for some variable-factor relationships while at the same time increase proportionally for other variable-factor relationships (in an orthogonal rotation). While the distance between variables and factors has changed, the relative distance between variables has not changed. In other words, the total common variance between variables estimated in the initial factor solution is conserved.
Browne (2001) describes how in the early days of factor analysis rotations were conducted by hand. This task took weeks, and adjustments to the factor axes was influenced by the configuration of points as well as subjective and substantive input from the researcher. However, the advent of computers and statistical software capable of approximating simple structure changed this process. Deliberate and informed rotational adjustments were automated by complexity function minimization.

There are a variety of complexity functions used to guide the rotation process (Bandalos, 2018; Browne, 2001). In general, the complexity function uses the factor loadings as inputs, and results in a summary measure of factor complexity, variable complexity, or a combination of the two. Factor complexity is the degree to which variables load on a single factor. A rotation that results in a lower number of variables loading onto each factor decreases the factor complexity. Variable complexity is the degree to which a single variable loads onto multiple factors. A rotation that minimizes the loading of a variable onto more than one factor reduces variable complexity.

Automated Rotations

The basic operation of the factor complexity function is to add the products of elements of rows together in one weighted term, and then add the products of the elements of columns together in another weighted term. Both terms added together measure the degree of complexity in terms of rows and columns based on Thurstone (1947)’s definition of simple structure.

The basic formula for the Crawford-Ferguson family of rotations (Crawford & Ferguson, 1970 as cited in Browne, 2001) is:

\[
c(s) = (1 - K) \sum_{i=1}^{f} \sum_{m=1}^{f} \sum_{i \neq m} \lambda_{im}^2 \lambda_{ii}^2 + K \sum_{m=1}^{f} \sum_{i=1}^{v} \sum_{k \neq j} \lambda_{im}^2 \lambda_{km}^2 \tag{2}
\]
where \( v \) is the number of items, \( f \) is the number of factors, and \( K \) is a constant that is used to differentiate between different versions of the Crawford-Ferguson rotation. The term left of the addition symbol is the total row (item) complexity, and the term right of the addition symbol is the total column (factor) complexity.

Using data from Table 1 as an example, the first term that totals the item complexity is calculated as follows.

\[
c(s) = \sum_{i=1}^{v} \sum_{m=1}^{f} \sum_{l \neq m}^{f} \lambda_{im}^2 \lambda_{il}^2
\]

\[
= .9^2(.05^2) + .8^2(-.1^2) + .3^2(.6^2) + .4^2(.6^2)
\]

\[
= 0.002 + 0.65 + 0.032 + 0.058
\]

\[
= .08
\]

**Table 1**

*Example Factor Loadings Table*

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.9</td>
<td>.05</td>
</tr>
<tr>
<td>2</td>
<td>.8</td>
<td>-.1</td>
</tr>
<tr>
<td>3</td>
<td>.3</td>
<td>.6</td>
</tr>
<tr>
<td>4</td>
<td>.4</td>
<td>.6</td>
</tr>
</tbody>
</table>

Pairs of factor loadings where one element is close to zero results in a smaller product than pairs of loadings where neither element is close to zero. Items where there is only one large factor loading and the rest of are smaller is less complex. Items where there are multiple large factor loadings are more complex. The second term (factor complexity) is calculated in much the same way:
\[ c(s) = \sum_{m=1}^{f} \sum_{i=1}^{v} \sum_{j=1, j \neq k}^{v} \lambda_{im}^2 \lambda_{km}^2 \]

\[ = .9^2(.8^2) + .9^2(.3^2) + .9^2(.4^2) + .8^2(.3^2) + .8^2(.4^2) \ldots \]

\[ = 1.03 \]

If \( k = 1 \) then:

\[ c(s) = (1 - 1)(.08) + (1)(1.03) \]

\[ c(s) = 1.03 \]

and the factor complexity is prioritized. If \( K = 0 \), then \( c(s) = .08 \), and the item complexity is prioritized. If \( 0 < K < 1 \), then the complexity function will reflect a weighted sum. Software rotates the factor loadings to minimize the complexity function, resulting in a factor loading matrix closer to simple structure. Notice that the value of \( K \) will determine what simple structure means and result in different factor loading matrices.

**Target Rotation**

While there are several different complexity functions based on minimizing item and factor complexity, Browne (2001) described a less data-driven approach called target rotation. Rather than the software minimizing weighted sums of complete columns and complete rows, the researcher specifies target values for a selection of factor pattern coefficients through an \( i \) by \( f \) target matrix \( B \).

\[
B = \begin{pmatrix}
? & 0 \\
? & 0 \\
0 & ? \\
0 & ? 
\end{pmatrix}
\]
In this 4 x 2 example target matrix, the 0 elements represent target values, and the question marks represent unspecified elements. The complexity function used to achieve target rotation uses information from the factor loading matrix and information from the target matrix (Browne, 2001):

\[ f(L) = \sum_{j=1}^{f} \sum_{i \in f} (\lambda_{ij} - b_{ij})^2 \] (3)

Let \( \lambda_{ij} \) represent factor loadings that correspond to \( b_{ij} \) values of 0 in the \( B \) target matrix. Using the factor loadings in Table 1:

\[
\begin{align*}
&= (.05 - 0)^2 + (-.1 - 0)^2 + (.05 - 0)^2 + (.3 - 0)^2 + (.4 - 0)^2 \\
&= .003 + .01 + .003 + .09 + .16 \\
&= .266
\end{align*}
\]

If the target loading is specified to be 0, then this equation amounts to the sum of squared loadings of the elements that are to be minimized. Rotating the matrix in a way that reduces .266 will result in a factor loading matrix with minimized loadings in specified element locations and maximized loadings in unspecified element locations.

This sum of squared differences allows for a partially specified factor pattern matrix, similar to a confirmatory factor analysis with two key differences. First, target rotation gets the specified loadings as close to 0 as possible without constraining them to be 0 (the relative relationships between loadings are preserved). This means that if the factor loading for the population is larger than zero, a misspecified model that assumes a zero loading can still be detected. Second, the relationships between loadings and correlations between factors derived prior to target rotation provide optimal fit before and after target rotation.
Browne (2001) suggests that iterations of target rotation can be used in a “non-mechanical exploratory process, guided by human judgment” (p. 123). After the initial EFA with target rotation was conducted, the researcher could identify loadings that were misspecified as zero and change the corresponding targets to be unspecified. Similarly, unspecified elements that approached zero could be changed to 0 in the target matrix. The factor analysis would then be run again, and additional adjustments made until the researcher was satisfied. Such an operation would be change this from a confirmatory approach to an exploratory approach.

Using target rotation increases the degree of specification of the model, and thus the complexity of the hypothesis. In a CFA model, researchers hypothesize exact values for non-target loadings (secondary loadings), while in EFA with target rotation, researchers hypothesize approximate values for non-target loadings. Model fit in a CFA can be evaluated relatively objectively using a limited number of global fit statistics. Model fit in an EFA with target rotation is a matter of both global fit and local fit. The researcher must decide when a factor loading deviates too far from the approximate 0 specified in the rotation. In this sense, it may be more difficult to evaluate whether the results obtained from an EFA with target rotation supports the acceptance or rejection of the hypothesized model.

**Exploratory Structural Equation Modeling**

**Definition and Features**

ESEM is a relatively new factor analytic approach (Asparouhov & Muthén, 2009), only available in Mplus software (beginning with Version 5.21). Some researchers have argued that ESEM combines the strengths of both EFA and CFA, and is preferred when investigating complex psychological constructs (Arias et al., n.d.; Asparouhov & Muthén, 2009; Boffò et al., 2012; Erreygers & Spooren, 2017; Marsh et al., 2014; Perry et al., 2015).
In their paper introducing ESEM, Asparouhov and Muthén (2009) argued that the use of CFA has led to several undesirable outcomes. First, they suggest that constraining small cross-loadings (even substantive ones) to zero contributes to misfit, and forces variance to be expressed in other parts of the model, inflating correlations between factors, and resulting in a model that is oversimplified. Secondly, when models lack fit in a CFA study, researchers who have started out in a hypothesis testing, deductive, theory-driven pursuit are forced to consult data-driven diagnostic tools like modification indices. A number of researchers suggest that extensive use of modification indices lead to incorrect models (MacCallum, 1992, as cited in Perry et al., 2015), capitalize on chance (MacCallum 1992), add atheoretical complexity (Schellenberg et al., 2014), and are overly responsive to the specific sample (Boffo et al., 2012; Reis, 2019), all of which threaten the replicability of the model. Marsh et al. (2014) argued that the strictness of CFA is primarily responsible for the failed replication of EFA findings in the Big Five personality research literature. As McCrae et al. (1996) argued, “There is no scientific utility in discovering the correct number of factors if I cannot reliably identify the factors because they fail to replicate from sample to sample” (p. 556).

Others have noted that the inflated correlations among factors that occurs if trivial and non-trivial cross-loadings are constrained to 0 may result in multicollinearity in SEM (Asparouhov & Muthén, 2009; Joshanloo, 2016a, 2016b; Marsh et al., 2010; Marsh, Nagengast, & Morin, 2013; Perry et al., 2015; Reis, 2019). The degree to which factors overlap may unduly influence the estimation of factor relationships with other variables in MIMIC models. Referring to the previous example, if researchers were attempting to differentiate between creativity and critical thinking as forms of intelligence, relating these two factors or constructs to different outcomes (grades in art class vs grades in science) might be an important element in a program
of validation. However, if CFA resulted in inflated correlations between these two factors, their relationship to other variables may be indistinguishable.

After identifying some of the main issues that result from using CFA on complex psychological data, Asparouhov and Muthén (2009) outlined how ESEM may serve as a reasonable alternative. In many ways ESEM is an extension of traditional EFA, as SEM was an extension of CFA. Currently (2021), Mplus is the only commercial software that explicitly facilitates ESEM. The Mplus User Guide (Muthén & Muthén, 2017) has examples of ESEM code (Examples 5.25-5.30, pp. 95-105) as well as EFA code (Example 4.1, p. 45). The differences are instructive. In Figure 1, I generated simple variants of these examples to compare different models by their description, code, and path diagram.

ESEM is an umbrella term which could describe several configurations. ESEM could be a single exploratory factor analysis, associating a vector of items with a vector of factors using data-driven rotations (in Mplus the default is the geomin rotation). Alternatively, specifying the use of target rotation invokes a confirmatory variant of ESEM, where EFA factor loadings are specified but not constrained to be zero on non-target factors (Rogoza et al., 2018). In other models, ESEM could include one EFA factor and relate it to one CFA factor. Additionally, though not shown in the table, ESEMs can test for measurement invariance, include independent variables in MIMIC models, or correlate error terms in longitudinal studies (Marsh et al., 2014; Muthén & Muthén, 2017).

In contrast to traditional coding methods for EFA (identify a rotation and number of factors), ESEM coding and output approximates the coding and output of CFA and SEM. Factors are identified by name, and factor to item relationships, factor to factor relationships, and error correlations are often specified a priori. This is particularly true when target rotation is used.
Figure 1

Comparison of EFA and ESEM Descriptions, Code, and Path Diagrams

<table>
<thead>
<tr>
<th>Model</th>
<th>Code</th>
<th>Path Diagram</th>
</tr>
</thead>
<tbody>
<tr>
<td>EFA:</td>
<td>Identify a range of models. In this case, only one model with 2 factors.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ANALYSIS: TYPE = EFA 2;</td>
<td></td>
</tr>
<tr>
<td>ESEM:</td>
<td>Name and identify factors 1 and 2 as EFA factors with (*1). Factor 3 is a CFA factor.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MODEL: f1-f2 BY x1-x15 (*1); f3 BY x16-x18;</td>
<td></td>
</tr>
<tr>
<td>ESEM:</td>
<td>Two EFA factors with theory imposed target rotations (~0), and one CFA factor.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ANALYSIS: ROTATION = TARGET; MODEL: f1 BY x1-x7 x8-x15<del>0 (*1); f2 BY x8-x15 x1-x7</del>0 (*1); f3 BY x16-x18;</td>
<td></td>
</tr>
<tr>
<td>ESEM:</td>
<td>Two blocks or sets of EFA factors with theory imposed target rotations.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ANALYSIS: ROTATION = TARGET; MODEL: f1 BY x1-x4 x5-x9<del>0 (*s1); f2 BY x5-x9 x1-x4</del>0 (*s1); f3 BY x10-x13 x14-x18<del>0 (*s2); f4 BY x14-x18 x10-x13</del>0 (*s2);</td>
<td></td>
</tr>
</tbody>
</table>
Similarly, ESEM tables produced by Mplus appear almost identical CFA or SEM tables, with item loadings or factor correlations appearing on an indented row under each specified factor, along with their corresponding statistical significance. In Mplus, ESEM is a an EFA, but an EFA that has been brought into the realm of structural equation modeling.

**Measuring Psychological Constructs**

There are several reasons that ESEM might be particularly beneficial at the initial stages of scale development for complex and closely related constructs.

Usually, even at the beginning of scale development, a theory driven model exists. Item generation is expensive in terms of time, energy, and often money. Scale developers will often have used a table of specification to assist in the organizing and drafting of items. Even if the scale developers do not make an explicit path diagram, this table of specifications serves as the model that is being tested (assuming the table differentiates between constructs and relates specific items to specific constructs). In their overview of factor analysis, Bandalos and Finney (2019) argue that unless the items in a scale were selected at random, “then some theory, however rudimentary . . . should be explicated to the extent possible” (p. 112). A theory, no matter how fledgling, supports a tentative hypothesis-testing approach.

When there is a clear model to be tested, compared with EFA, ESEM with target rotation is a much more effective confirmatory tool. Where EFA may mute some of the specificity of the model in favor of data-driven model fit, ESEM with target rotation allows the model to be closely approximated without a reduction in fit. The added model specification contributes to the validity of a statistical test. In a traditional EFA, the model being tested is not the model that was used to generate the items for the scale. Additionally, the constraints of a model are particularly important when testing psychological constructs that may be closely related. As described
earlier, data driven approaches to model generation may result in nuisance factors, require the researcher to identify a factor based on item content, and may result in confusing models when loadings are rotated to minimize a-theoretical factor and item complexity. The imposition of a model on the rotation of factor loadings may bring additional strength to statistical inferences and stability to the model. In their simulation, Asparouhov and Muthén (2009) found that compared to geomin, target rotation performed well, with negligible bias and better coverage. Finally, unlike traditional EFA, ESEM can easily relate the results of the scale to other theoretically important variables (MIMIC models) as a form of additional validity testing.

While Marsh et al.’s (2014) seminal paper emphasized the use of ESEM as a confirmatory tool, they also acknowledged the potential for ESEM to be used in an exploratory manner. Compared to CFA, ESEM with target rotation provides more useful diagnostic insights. It is common at the beginning of scale development for items to behave differently than expected. Like CFA, ESEM with target rotation allows for model specification, but unlike CFA it also results in EFA like diagnostics. In other words, in the likely case that a beginning model does not fit well, ESEM with target rotation allows items loadings that do not approximate zero to be identified, which is more useful in the process of model-specification and theory generation than modification indices commonly used in CFA (Browne, 2001). Additionally, target rotation has been shown to perform well compared to CFA. In a follow up simulation study to Asparouhov and Muthén’s (2009), ESEM and target rotation were compared with CFA, and ESEM not only returned less inflated correlations among factors, but effectively estimated the population parameters (Marsh, Lüdtke, et al., 2013).

However, there are also several limitations. Compared to a CFA, the additional number of parameters being estimated in ESEM is substantial. All things being equal, simpler statistical
models are preferred. Even if there is greater model fit in an ESEM, the additional complexity of
the model must be weighed against the model fit. Additionally, increasing the number of
parameters complicates the interpretation of the results, especially by practitioners that might use
the scale. The need to account for substantive cross-loading between items of different factors
detracts from the usability of the scale. Finally, as of the current version of Mplus (8.1), there is
no simple way to test higher order factor structures like those that are implied by the model used
to generate items for the SEWS Beta Form A. In such cases, an exploratory structural equation
model inside a confirmatory factor analysis (EwC) is used (Marsh et al., 2014). The ESEM factor
loadings and factor correlations are obtained for the first-order factors. The factor loadings and
correlations are then specified as starting values in a model that contains second-order factors
(Marsh et al., 2014).

One alternative proposed by Marsh et al. (2014), is to use both ESEM and CFA and
compare the results. This might be a particularly meaningful approach at the beginning of scale
development because it provides the scale developer insights that ESEM or CFA cannot provide
separately. In the context of the SEWS Beta Form A, while it is possible that items will cross-
load in non-trivial and perhaps even substantively meaningful ways, it is not clear which items
will cross-load, how extreme the cross-load will be, or how much such cross-loading will affect
correlations between factors if they are constrained to 0. ESEM could give insight into the degree
to which items cross-load, and the difference between correlations between factors in the CFA
model and the ESEM can indicate to what degree CFA correlations are inflated by the item-level
constraints.

Asparouhov and Muthén (2009) warn that ESEM is not meant to replace CFA. Rather
ESEM was designed in response to the reality that real life data can be messy, and that the
simple structure of a CFA may be inappropriate in some circumstances. ESEM was meant to be
a tool in the factor analysis toolbox. Some researchers have used ESEM in conjunction with CFA
as a measure of robustness (Chung et al., 2016; Hukkelberg et al., 2018; Neff et al., 2019;
Rathwell & Young, 2016). When ESEM and CFA models result in similar loadings and factor
correlations, the researcher can have more confidence in the parsimony of the CFA. When they
are different, the researcher can use the ESEM findings to identify problematic item loadings.
Marsh et al. (2014) recommended that ESEM be used with CFA, but, when possible, to prefer
the CFA results.

Testing Rival Models

When a confirmatory approach of factor analysis is being used, testing rival models can
lead to deeper insight into the dimensionality underlying the covariance between observed
variables. Bandalos and Finney (2019) remind researchers that scientists are not just responsible
for seeking out evidence that their model is false, but they are also responsible for considering
evidence of equally plausible rival models. This implies that the alternative models under
consideration provide distinct interpretations of the data, and that they fit the data equally well.

Marsh et al. (2014) claims that ESEM is a more generalized case of the CFA, allowing
researchers to compare these models using approaches reserved for hierarchically nested models.
Nested models must have the same items but will differ in the number of parameters that are
estimated. If relaxing or constraining parameters can transform one model into the other, then it
is likely the models are nested. In the case of a generic measurement ESEM and CFA, both the
relaxing and constraining of parameters that relate the two is relatively straightforward. Any
factor-loading for cross-loading items would be constrained to 0 to produce the CFA model.
Nested models can be compared statistically using likelihood ratios and the chi-square difference test (Bandalos & Finney, 2019). This test must be adjusted from the typical chi-square difference test to consider parameter boundaries (e.g., correlation coefficients can only range from −1 to 1; Bandalos, 2018). This is particularly important in the common factor model, latent growth curve models, and autoregressive models, where the traditional distribution of the test statistic has tails that are heavier than the true distribution, resulting in a failure to detect differences when they exist (Stoel et al., 2006). Instead, critical values must be obtained from the chi-bar distribution (Bandalos, 2018).

However, as noted by many authors, models should only be compared if they have acceptable absolute fit (global fit, high factor loadings, appropriate correlation between factors values; Bandalos & Finney, 2019; Brown, 2015). Also, there is no use comparing a well-fitting model with a poorly fitting model because a poorly fitting model is of little use. At the end of their chapter on factor analysis, Bandalos and Finney (2019) emphasize that there is no true model, but a model that has been chosen based on supporting evidence. They also remind readers that the fit of the chosen model, and the reasoning for rejecting equally viable rival models should be given equal weight. In other words, researchers should not focus solely on the lack of evidence to reject their chosen model, but they should also spend time reporting the evidence that was used to reject alternative models. This might be especially important when comparing CFA and ESEM results. Booth and Hughes (2014) found that in many cases, it was difficult to justify the increased complexity suggested by the ESEM when compared with its rival CFA model.
Reliability

Problems With Cronbach’s Alpha

Reliability is often described in the psychometric literature as a degree of consistency, trustworthiness, dependability (Bandalos, 2018, p. 155), or repeatability (Cortina, 1993). It could be the degree to which results would be replicated across testing occasions, or across parallel forms of a tests, or the degree to which scores are consistent within a single sample of scores (Bandalos, 2018). Reliability is important because measurement always includes error. The degree to which scores are inconsistent indicates the level of error included in the scores.

Cronbach’s alpha (Cronbach, 1951), or coefficient alpha, is used almost exclusively in the literature as a coefficient of reliability, especially internal consistency reliability, and has become synonymous with reliability in general (Cortina, 1993). Based on work by Spearman and Brown, Kuder and Richardson, and Guttman (as cited in Cho & Kim, 2015 and Sijtsma, 2009a), Cronbach generalized the Kuder-Richardson formula used to estimate reliability of dichotomously scored items, allowing it to be used with polytomous scored items. Because of this contribution, in classical test theory, the reliability of a composite score (i.e., the total or average) can be estimated using a single sample without multiple testing occasions, or the use of parallel tests (Cronbach, 1951). It’s comparative ease of use, popularity in the literature, and the degree to which it is readily available on statistical software such as SPSS all contribute to its continued use in the field, despite long standing objections from psychometricians (Sijtsma, 2009a).

Researchers have been publishing about the dangers of using Cronbach alpha since as early as Cortina’s 1993 critique. Alpha assumes that the items load on the factor with relatively similar magnitudes (i.e., essential tau-equivalence), and that they are unidimensional (Peters,
Green and Yang (2009) argue that it is unlikely that any scale will be purely
unidimensional, and even in small scales that measure a narrow construct, it is unlikely to
achieve tau-equivalence. When these assumptions are not met, coefficient alpha is the lower
bound of the reliability, meaning it is an underestimation of the true level of interrelatedness
(Cho & Kim, 2015). A lower bound of 0.8 suggests that the true reliability of a composite lands
somewhere between 0.8 and 1. However, Green and Yang (2009) also suggest alpha could be
moderately robust to this violation if the average factor loading is above a .6, and there are no
extreme differences between factor loadings.

A third underlying assumption is that errors are uncorrelated. Green and Yang (2009)
have shown that error correlations will usually result in an artificially inflated coefficient alpha,
overpromising and underdelivering. Alternatively, depending on the deviation from tau-
equivalence relative to the correlation in error terms, coefficient alpha could overestimate or
underestimate reliability (Cho & Kim, 2015). In either case, the bias tends to be much more
pronounced when errors are correlated, meaning that coefficient alpha is less robust to this
violation (Green & Yang, 2009). What is more troubling, is that sources of such variance are not
rare and could include (a) response sets resulting from item wording, (b) order effects, or (c)
stimulus material (Green & Yang, 2009; Sijtsma, 2009b).

Even when these assumptions are met, it is unclear how to interpret coefficient alpha.
While it is commonly regarded as a measure of internal consistency, Cho and Kim (2015) make
a detailed argument that the three possible interpretations of what internal consistency means
probably do not apply to coefficient alpha. Coefficient alpha is not a measure of internal
consistency as homogeneity or unidimensionality. Using a table of four correlation matrices, Cho
and Kim outlined why this is the case. Correlation matrix A reflected a single factor, with
correlations of .3 on all the off diagonals, and correlation matrix D reflects a single general factor with two group factors. Coefficient alpha for correlation matrix A was 0.77, while the coefficient alpha for correlation matrix D was 0.87. In other words, the matrix that was unidimensional resulted in a lower coefficient alpha than the matrix which was multidimensional. Nor does internal consistency mean general factor saturation. In Cho and Kim’s example, a matrix reflecting two factors result in the same reliability coefficient as one weak general factor. Similar demonstrations and arguments were made by Schmitt (1996).

Although it is more widely accepted, internal consistency is not an absolute stand in for interrelatedness of the items because coefficient alpha is a function of both interrelatedness and the number of items in the set. A set of items that have interitem correlations as low as .1 can still achieve a .7 coefficient alpha with only 21 items (Cho & Kim, 2015). Despite arguing that internal consistency does not exclusively indicate relatedness, Cho and Kim (2015) suggest that the term *item interrelatedness* is less ambiguous than internal consistency reliability.

**McDonald's Omega**

When basic assumptions of tau-equivalence and uncorrelated error terms cannot be met, Cho and Kim (2015) recommend using a model-based reliability approach. Specifically, they recommend using the multidimensional version McDonald’s omega when the data fits a multidimensional model well, and the stratified-alpha if the data is multidimensional but does not fit a multidimensional model well.

Model-based reliability is a more appropriate estimate of internal consistency because it is calculated using the correlations or covariances obtained from the items (Revelle & Condon, 2019). While coefficient alpha can be used under specific model conditions (i.e., unidimensionality, tau-equivalence, and uncorrelated errors), there are families of reliability
coefficients for a variety of model configurations. For example, the family of McDonald’s omega coefficient includes a generic omega total ($\omega_t$), omega hierarchical ($\omega_h$), and an omega bifactor solution ($\omega_g$; Revelle & Condon, 2019; Zinbarg et al., 2005).

Model-based reliability estimates take advantage of the fact that factor loadings represent common variance partitioned out from unique item variance or error variance. The formula for the omega coefficient ($\omega$; Bandalos, 2018, p. 395) uses the factor loadings ($\lambda_i$) and residuals obtained from a CFA:

$$\omega = \frac{(\sum \lambda_i)^2}{(\sum \lambda_i)^2 + (\sum \theta_{\delta_i}^2)}$$ (4)

The numerator represents the sum of the variance in each item that is explained by the factor, a close approximation of the true score. The denominator represents this approximate true score variance added the unexplained variance, or the total variance. Overall, it is a ratio, of true score variance over total variance. This results in a proportion, where $0 < \omega < 1$. The larger the coefficient, the more the consistent the scores were.

**Summary**

Measuring SEL is an important step in providing a more holistic education for students. There are many frameworks that identify a variety of skills, attitudes and knowledge that are important for successful relationships and high academic achievement. The framework that ASD has developed identifies four main constructs (Safety, Resilience, Self-mastery, Confidence) with four sub constructs (Resilience: Bounce-back, Resilience: Perseverance, Self-mastery: Self-awareness, Self-mastery: Self-management). A student's sense of Safety at school moderates their learning. A student's ability to Bounce-back and Persevere may be indicators of their overall ability to face challenges with Resilience. A student’s awareness and regulation of emotions directly contributes to or detracts from their relationships with others, and indirectly
affects their academic achievement. Finally, how Confident a student feels about their life indicates their sense of optimism about the future and their abilities to face challenges.

In the process of developing scales for complex psychological constructs, new techniques such as ESEM may provide meaningful insight into the relationships between indicators and latent variables. Using target rotation, researchers can specify a model without constraining factor loadings on non-target factors to zero. Used in conjunction with CFA, ESEM may provide insights that will be beneficial to developing scales and models that are more replicable. I hypothesized that comparing results from ESEM and CFA approaches would be beneficial in assessing the factor structure of the SEWS Beta Form A. Additionally, model-based reliability estimates like a variant of McDonald’s omega may be a more appropriate way of estimating the consistency of the internal structure.
CHAPTER 3

Method

Prior Steps in Scale Development

While the purpose of this study was to evaluate the initial pool of items generated for the SEWS Beta Form A, a brief overview of the preceding steps will be given. ASD’s Vision for Learning incorporates academic and non-academic outcomes into a holistic framework inspired by models like the 6 C’s, 21st Century Learning, and CASEL. The non-academic outcomes are comprised of conditions (Safety, Confidence, Connection) and dispositions or skills (Self-mastery, Resilience, Compassion, Respect). Significant effort went into defining and prioritizing these constructs, involving large groups of teachers and principals from multiple schools and school levels as well as district employees.

In collaboration with ASD’s Social and Emotional Well-being Team, I explored existing scales. This included a review of the commercial scales developed by companies like Panorama and Pearson, as well as a search of databases such as the Mental Measurement Yearbook with Tests in Print, APA PsycTests, and Health and Psychosocial Instruments. While this exploration was not exhaustive, pricing as well as trends in the free-to-use scales led to the decision to develop a new scale. Free-to-use scales that were rejected had three common issues that highlight some of the challenges that districts face in adopting a high-quality measure.

First, some scales probed for sensitive information. For example, 9 out of 10 items that are intended to measure perceptions of safety in the MDS3 Student Survey deal with violence, drug use, and alcohol use (Bradshaw et al., 2014). The MDS3 was developed by the John Hopkins Center for Youth Violence Prevention, with focus group input by students, administrators, and district personnel. Though the MDS3 was created and refined using rigorous
methods of scale development to be used by students in a public education setting, these types of items can raise the bar for what could be considered valid consent. Such a measure would have required ASD to obtain opt-in permissions from parents, resulting in increased classroom interruption, and additional political and social hurdles.

Second, the items in some scales did not align with ASD’s definitions of constructs or target audience. As an example, one self-control scale had items that were too general, referring to the use of money, preparing for the future, or concern for the long run (Cochran, 2016). These items did not reflect the school context or the shorter-term sense of endurance and impulse control that ASD defined as Perseverance. This was a common issue. Another common issue was the alignment of sources and targets of data. For example, the search term “Respect, student” in the Mental Measurement Yearbook with Tests in Print and APA PsycTests resulted in 35 hits between 1960 and 2018. Only five were relevant to an educational setting pre-university, all of which were either too specific (respect for subgroups) or were collecting data from the wrong source about the wrong target (teacher’s perception of student respect, or students’ perception of teacher respect). In contrast, ASD was interested in a self-report on self-perception measure.

Third, scales sometimes used words or phrasing that might be distracting or confusing for children and teenagers. The Brief Self-Control Scale—Alternate Version (BSCS) contains items such as “Pleasure and fun sometimes keep me from getting work done” or “I am good at resisting temptation” (Kirby et al., 1999). Similarly, the Self-Control Scale uses words such as “deliberately” or “neglect”, which might not be well understood at the elementary level (Jeong et al., 2016). Finally, some scales like the Eysenck Impulsivity Scale – Modified Version contained
a mix of or all negatively valenced items, whereas ASD was interested in positively valenced items only (Neumann et al., 2010).

The decision to adopt an instrument or develop one is not an easy decision to make. The widely used text on survey development *Internet, phone, mail, and mixed-mode surveys: the tailored design method* by Dillman et al. (2014) highlights the difficulty of generating items that are easily understood, specific, and meaningful. However, while Dillman et al. (2014) suggest reviewing other scales, they warn that researchers must be particularly careful that the adopted scale align with the intended purposes and scope of the research questions – scores obtained by a scale that are valid and reliable in one context is not a guarantee of validity and reliability in another context. Having failed to find a scale, or sets of scales, that satisfied the purposes and scope defined by the Social and Emotional Well-Being Team at ASD, they opted to develop their own.

Based on the definitions for each construct, a pool of items was generated, evaluated, and revised. This involved creating a matrix of possible contexts for each skill and condition, and then brainstorming observable behaviors or pertinent feelings and perceptions within those contexts. Anticipating that our initial trials with these items would result in some items functioning poorly, we chose to generate a surplus of two to three items per sub-construct.

During phases of this generation, we sought evidence of construct validity by involving participation from outside the initial team tasked with writing the items. In the first activity, teachers were asked to sort items into categories without having been told which construct they had been generated to measure. The results of this activity were informative, and items were revised. In the second activity, third graders, and a spread of junior high school and high school students were recruited to participate in a think-aloud study. The items were reviewed and a
portion of items that were most likely to be misunderstood were identified. Parents were
contacted and consent was received beforehand. In the think-aloud activity, students read
approximately 10 items from the flagged item pool out loud and identified how they would respond and why. This study revealed some nuance in how students were understanding the items. For example, we discovered that elementary school students interpreted “I feel like my teacher understands me” to mean that their teacher understands what they say grammatically or that their teacher understands them while wearing a mask, whereas the item had been written to measure the degree to which students felt that their teachers understood them emotionally and intellectually. Again, revisions were made based on these and other findings. The items were then reviewed by legal counsel and the Student Educational Equity team. Additional revisions were suggested and implemented.

The final pool consisted of 81 items intended to measure 14 first-order constructs related to the conditions and dispositions identified in ASD’s Vision for Learning framework. Because this was a pool much larger than was reasonable to have elementary school students respond to at one time, and because of scheduling issues and other practical complications with recruiting junior high school and high school participation, ASD chose to create two forms. The Beta Form A consisted of the 38 items and constructs described in this paper. The Beta Form B consisted of 43 items intended to measure other constructs (e.g., Compassion, Respect, Connection, Equity, etc.). Because we anticipated that some constructs would be highly correlated with other constructs, in the item reducing step of the scale development we wanted these possible converging factors to be in the same form. By being in the same form, we would be able to reduce the possibility that these factors converged by eliminating items that cross-loaded highly between them. Both forms were planned to be administered in the Spring of 2021, but due to
further complications only the Beta Form A was administered. Further development of the Beta Form B is planned for the Fall of 2021.

The remainder of this paper is dedicated to the pilot study conducted on the Beta Form A.

Participants

The student population of ASD consists of majority White students, from diverse socio-economic backgrounds. Three schools were selected with the intent to reflect the socio-economic diversity. These three schools included one elementary school, one junior high school, and one high school. Complications with obtaining permissions, additional stresses caused by the Covid-19 pandemic, and concerns from school board members resulted in a much smaller sample size than I previously anticipated.

Of the total 461 student participants, 187 were from a junior high school (7th = 64, 8th = 73, 9th = 44) and 274 were from a high school (10th = 167, 11th = 52, 12th = 55, NA = 6). Additionally, there was an even spread of male and female participants (Male = 228, Female = 214, Other = 12, NA = 7). Most of the sample consisted of White students (White = 397, non-White = 54, NA = 10).

Measures

The SEWS Beta Form A is a 38-item questionnaire designed to measure the following six constructs: (a) Safety (7 items; e.g., “I feel safe to make mistakes at school”), (b) Confidence (7 items; e.g., “I believe I will succeed in life”), (c) Bounce-back (6 items; e.g., “When life is hard, I stay strong”), (d) Perseverance (6 items; e.g., “I stay motivated during the school day”), (e) Self-management (6 items; e.g., “When I'm upset, I stop and think before I do something”), and (f) Self-awareness (6 items; e.g., “I can name my feelings”). All items are positively valenced.
Two sets of response continuums were selected. In both cases, the response continuums were unbalanced, consisting of two response options that indicate low frequency or low agreement, and three response options indicating high frequency or high agreement. Students responded to well-being constructs or conditions constructs (i.e., safety and confidence) using a 5-point rating scale ranging from 1 (disagree) to 5 (strongly agree). Students respond to skill or disposition constructs (i.e., Bounce-back, Perseverance, Self-management, and Self-awareness) using a 5-point rating scale ranging from 1 (almost never) to 5 (always).

This unbalanced continuum was selected for two reasons. First, ASD is interested in measuring progress. By allowing more nuance on the positive end of the continuum, the instrument would be capable of collecting information about subtle changes that might normally have been obscured by a more limited set of response options. Second, when performing cognitive testing of the items on students using a balanced continuum, one junior high student indicated “almost always” but voiced that he would rather have indicated “always.” The reason for not including the “strongly disagree” and “never” options was to avoid overwhelming younger students with six response categories.

Each construct was separated by a page break in Qualtrics, meaning that there were never more than 7 items presented at a time. Each page was labeled with the construct name (e.g., “Safety,” “Bounce-back”).

**Procedures**

Consent forms in English and Spanish were sent to parents and guardians on March 17, 2021, at a junior high school in Alpine School District (ASD). Guardians then responded positively or negatively through the ASD data management services Skyward. Students that obtained permission were sent a link to the SEWS Beta Form A survey using Qualtrics.
However, due to unforeseen difficulties in the process of collecting permissions, only 187 junior high students were able to participate. To alleviate some of these difficulties, ASD switched to an opt-out format, in which students and parents were informed of the survey, its purpose, and its current stage of development, and guardians were allowed to opt their student out of the survey before it was administered. This allowed us to collect an additional 274 high school student responses.

For the purposes of this dissertation, my role was limited to analyzing the data collected by ASD. To publish my findings, I sought and obtained permission through the Brigham Young University’s Institutional Review Board process and subsequently received permission through ADS’s review board. Documentation obtained from both organizations can be found in the appendix.

**Analysis**

*Missing Data*

Approximately 3% of the data was missing. After examining the cells with missing data, I discovered that 93% of the missing data came from 13 students at the junior high level who did not complete most of the scale, 9 of whom did not complete any of the items. These 9 students were subsequently dropped from the analysis. As most of these students did not complete demographic items, it is unclear whether ethnicity, gender, or grade level contributed to the missing data. When regressing school membership on total missing observations per student, being a member of the junior high school was a statistically significant influence ($\beta = .19$).

The type of missing data influences the degree to which some estimators can compensate for missing data. For example, if the data are missing completely at random (MCAR; e.g., the internet dropped for a random set of students taking the survey), then listwise and pairwise
deletion will return unbiased parameter estimates and standard errors, though they will also result in a reduction of power due to the diminished sample size (Brown, 2015). However, if the outcome variables are related to some of the observed variables (e.g., demographics, previous test scores), then it would be considered missing at random (MAR). In the case of MAR or MCAR, when the data have a multivariate normal distribution, using the Direct ML results in unbiased parameters, standard errors, and test statistics. Because it uses all the data, Direct ML is also the most efficient estimator. Unfortunately, when using an estimator more appropriate for categorical data like the weighted least squares mean variance (WLSMV) estimator that Mplus employs, a pairwise deletion approach is used (Asparouhov & Muthén, 2010). Mplus reported dropping 9 rows of observations, estimating the factor analyses and correlations among factors using 452 rows.

This less-than-ideal handling of missing data may result in a loss of power, and possible bias in the estimates. However, due to the small percentage of missing data, I assumed the bias introduced would not have a compromising effect on my analysis.

**Factor Analyses**

Before running Model 1 as shown in Figure 2, a simplified version was evaluated. While Model 1 was the theorized model used to write the items, it includes two second-order factors. Specifically, Model 1 specified that the first-order factors called Bounce-back and Perseverance would load onto a second-order factor called Resilience, and that the first-order factors called Self-awareness and Self-management would load onto a second-order factor called Self-mastery. To assess whether this hierarchical configuration was justified, the correlations between first-order factors were examined prior to assessing the full hypothesized model. Thus, the initial CFA
Figure 2

Model 1
model and ESEM excluded the second-order factors, and kept the following six first-order factors: (a) Confidence, (b) Safety, (c) Perseverance, (d) Bounce-back, (e) Self-management, and (f) Self-awareness. It was expected that Bounce-back and Perseverance would correlate at .80 or above, and that Self-management and Self-awareness would also correlate highly. If the factors were highly related, a hierarchical configuration would be justified.

While the ESEM had excellent model fit based on Hu and Bentler’s (1999) recommendations (Tucker-Lewis Index, TLI > .96; Comparative Fit Index, CFI > .96; Standardized Root Mean-square Residual, SRMR < .09; Root Mean Square Error of Approximation, RMSEA < .06), all the correlations between factors were lower than .80. Additionally, the CFA model did not achieve acceptable fit (CFI = .94, TLI = .93, SRMR = .05, RMSEA = .07), and only one correlation between factors was above .80 (the correlation between Bounce-back and Perseverance was .82). Self-awareness and Self-management correlated at .65. The unexpected lack of convergence between factors, and lack of model fit did not provide the needed justification to evaluate the hierarchical configuration. Based on these initial findings, I did not evaluate the full model in Figure 2.

While our initial model failed to fit the data, relaxing some of the stringent hypothesis testing, and deductive reasoning constraints allowed me to investigate the dimensionality of the data more fully and explore alternative options. Specifically, the goal of this approach was to create a shortened version of the SEWS Beta Form A that (a) fit the data well, (b) preserved the theoretical meaning of the constructs, (c) differentiated between the constructs, (d) achieved measurement invariance in terms of gender and school level, and (e) assess the reliability of the scores for each construct. During the item generation process with district psychologists and
leaders, we had intentionally included more items than desired. The goal was to reduce this 38-item scale to a useful scale of approximately 20 items.

Working in collaboration with district psychologists, items were selected for omission based on statistical and theoretical grounds. In alignment with Bandalos and Finney’s (2019) guidance on item omission, we proceeded cautiously, omitting one to six at a time. We considered the content of the items and how the omission of certain items might affect the validity and meaning of the construct, as well as the relationships between other items and constructs. There was a danger that we would underrepresent the construct being measured by deleting items with substantive contributions for the sake of a better fitting model.

While preserving construct validity in terms of content was important, other statistical diagnostics were also considered. We examined the strength of primary loadings and the magnitude of cross-loadings, especially ones that were statistically significant and that corresponded with inflated correlations between factors. We also considered global fit statistics, the degree to which correlations between factors were inflated in the CFA model, modification indices, and Heywood cases. Intermittently, we would also examine measurement invariance of the models in terms of gender and school level, identifying items on which the slopes or intercepts were larger for one group than another.

We specifically looked for items that performed worse from multiple diagnostic sources, rather than using any single source to dictate the items to be dropped. For example, while modification indices tended to be less informative, one standardized item loading was larger than one (Heywood case), and an examination of the modification indices suggested allowing the error terms to covary between it and a similar item. In fact, on closer examination, the wording in the items was remarkably similar. These three sources of information provided meaningful
justification for omitting the item. After we had identified a set of items to drop, I re-ran the factor analysis, modeling both a CFA and ESEM with target rotation using the new set of items. In this way, we iteratively identified, dropped, and re-evaluated sets of items, and narrowed the scale to items that performed well and supported the intended purpose of the shorter version of the scale.

When higher-order models were considered, and because there were only two first-order factors for each second-order factor, factor variances for the second-order factors were set to 1 for identification purposes (Brown, 2015). However, the relationships between the items suggested an alternative model. The final model consisted of just the first-order factors.

CFA and ESEM were performed using Mplus software Version 8.5 (Muthén & Muthén, 2017) with the WLSMV estimator for categorical data. Based on the recommendations from Hu and Bentler (1999).

Currently, there are no cutoff recommendations for EFA using target rotation, so I used the cutoffs suggested for EFA with oblimin rotation. Factor loadings greater than .40 on their specified factor were salient or meaningful (Brown, 2015; Howard, 2016). Based on Howard’s (2016) recommendations, cross-loadings were defined as any factor loading on a non-specified factor that meets the following two criteria: (a) is greater than .30, and (b) is less than .20 higher or lower than the “primary and alternative factor loadings” (p. 55).

Reliability

Internal consistency reliability is the “degree to which responses are consistent across items within a scale” (Bandalos, 2018, p. 173). One commonly reported reliability coefficient is Cronbach’s alpha. For inferences about reliability to be valid when using Cronbach’s alpha, items must be unidimensional and essentially tau-equivalent (Dunn et al., 2014), meaning that all
factor loadings are the same, or if they are different, they are different by a constant. Also, errors cannot be correlated. Because these assumptions are usually not met, other reliability coefficients have been developed which are not based on satisfying these three assumptions. As tau-equivalence was not met in our final model of the SEW Beta Form A, I used the McDonald’s Omega coefficient, a model-based reliability coefficient used to estimate the reliability of each construct.

**Measurement Invariance Across Groups**

Finally, I tested for measurement invariance across gender and school level. For each grouping, I performed the following analysis. Using Mplus (Version 8.5), I tested for configural invariance which is an indicator that the items load onto the same factors for both groups. If there was evidence of configural invariance, I tested for metric invariance by constraining all the factor loadings to be the same across groups. If there was evidence of metric invariance, I tested for scalar invariance by constraining intercepts to be the same across groups. For metric and scalar invariance, a reduction in CFI of less than .01 was considered evidence of invariance (Cheung & Rensvold, 1999, as cited in Vandenberg & Lance, 2000). Because of the small cell size of the category “Other” ($n = 12$), these observations were omitted when I tested for measurement invariance across groups. A sensitivity test was run on the final model for the CFA and ESEM estimates using this smaller data set. However, because there were only three fewer observations, changes in loadings and correlations were minimal.
CHAPTER 4

Results

Shortened SEWS Beta Form A

Using the sequential and exploratory process of omitting items, a shorter version of the SEWS Beta Form A was identified.

The first set of iterations resulted in a shorter version of the scale with acceptable global fit and factor loadings that loaded highly on their target-factor. However, this first shortened version also resulted in undesirable correlations between factors. Specifically, the correlations between Bounce-back, Perseverance, and Confidence were above .80, resulting in a lack of discriminant validity among the first-order factors. When the second-order factors of Resilience and Self-mastery were included in the model, the second-order factors also had a correlation above .80.

With these relationships in mind, we conducted a second set of iterations, reintroducing all the original items. During the second attempt, we placed a higher priority on omitting items that cross-loaded meaningfully between Confidence, Perseverance, and Bounce-back. In consultation with ASD school psychologists, we also revised the intended model to omit the second-order factors of Resilience and Self-mastery. The final 23 items strongly loaded on their target factor, produced acceptable fit statistics, and returned correlations among factors lower than .80.

Global Fit

ESEM produced more acceptable fit indices because by definition this method allows items to cross-load on factors other than the targeted factor. However, both ESEM and CFA models resulted in acceptable fit, as shown in Table 2.
Table 2

Global Fit Statistics for the Shortened SEWS Beta Form A

<table>
<thead>
<tr>
<th>Model</th>
<th>CFI</th>
<th>TLI</th>
<th>SRMR</th>
<th>RMSEA</th>
<th>RMSEA 95% CIs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESEM</td>
<td>.99</td>
<td>.98</td>
<td>.01</td>
<td>.05</td>
<td>.03-.06</td>
</tr>
<tr>
<td>CFA</td>
<td>.97</td>
<td>.96</td>
<td>.03</td>
<td>.06</td>
<td>.06-.07</td>
</tr>
</tbody>
</table>

*Note.* ESEM = exploratory structural equation model; CFA = confirmatory factor analysis.

**Correlations Among Factors**

As shown in Table 3, the correlations among factors were lower in the ESEM model than in the CFA model. This was expected, because less inflated correlations between factors are a well-known result of ESEM (e.g., Booth & Hughes, 2014; Gomes & Gjikuria, 2017; Marsh et al., 2010; Marsh et al., 2014). The highest correlations between factors in the CFA model was .76 between Confidence and Bounce-back, and .75 between Bounce-back and Perseverance.

Table 3

CFA and ESEM Correlations Among Factors for the Shortened SEWS Beta Form A

<table>
<thead>
<tr>
<th>Factor</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Self-Aware</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Self-management</td>
<td>.66 (.59)</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Perseverance</td>
<td>.57 (.49)</td>
<td>.68 (.58)</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Bounce-back</td>
<td>.59 (.51)</td>
<td>.66 (.50)</td>
<td>.75 (.62)</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Confidence</td>
<td>.60 (.51)</td>
<td>.57 (.42)</td>
<td>.66 (.53)</td>
<td>.76 (.65)</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>6. Safety</td>
<td>.57 (.52)</td>
<td>.51 (.41)</td>
<td>.66 (.60)</td>
<td>.68 (.59)</td>
<td>.71 (.61)</td>
<td>–</td>
</tr>
</tbody>
</table>

*Note.* ESEM results are in parenthesis.
While these correlations are not above .80, they are still relatively large. In contrast, the correlations between factors in the ESEM model were .65 and .62 respectively, indicating that the inflated factor correlations in the independent cluster CFA model may be the result constraining the cross loadings to zero.

**Factor Loadings**

Table 4 displays the factor loadings for both the ESEM and CFA models. CFA factor loadings tended to be higher than corresponding loadings of items on their target factor in the ESEM model (e.g., SA_3 loaded on Self-awareness at .83 in the ESEM model and .85 in the CFA). While there are several statistically significant non-target loadings in the ESEM model, their loadings are all well below .40. For example, SM_2 loads onto its target factor Self-management at a .48, but also loads significantly on perseverance at .12 and Bounce-back at .14. However, the square of a factor loading indicates to what degree that factor explains the variance of that item, meaning that Perseverance ($12^2 = .01$) only explains 1% of the variance of SM_2, and Bounce-back ($14^2 = .02$) only explains 2% of the variance of SM_2. The only exception to this general finding was item BB_1, which loaded significantly on Perseverance rather than on Bounce-back as intended. The modification indices also indicated that switching BB_1 to Perseverance would result in better fit. After examining the content of the items, and in consultation with ASD personnel, we decided that switching BB_1 to load onto Perseverance was an acceptable change. Figure 3 shows the CFA model in a path diagram.

Overall, items in the shortened version of the SEWS Beta Form A load highly on their intended factor. We retained five Self-awareness items that had factor loadings ranging from .59 to .90. We retained four items for Self-management with factor loadings that ranged from .60 to .78. Perseverance consisted of four items with factor loadings that ranged from .76 to .92. In
Table 4

*ESEM and CFA Factor Loadings of the Shortened SEWS Beta Form A*

<table>
<thead>
<tr>
<th>Item ID</th>
<th>Item Text</th>
<th>ESEM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SA</td>
</tr>
<tr>
<td>Self-awareness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA_1</td>
<td>I notice when my feelings change (happy to sad, angry to calm).</td>
<td>.72</td>
</tr>
<tr>
<td></td>
<td>If I am feeling a strong emotion, I know why I'm feeling it (bad grade,</td>
<td>.81</td>
</tr>
<tr>
<td></td>
<td>made a new friend, or someone was mean).</td>
<td></td>
</tr>
<tr>
<td>SA_2</td>
<td>I know how I'm feeling (angry, sad, happy).</td>
<td>.83</td>
</tr>
<tr>
<td>SA_3</td>
<td>I can recognize and name my emotions.</td>
<td>.82</td>
</tr>
<tr>
<td>SA_4</td>
<td>I notice what my body does when I experience emotions (nervousness,</td>
<td>.73</td>
</tr>
<tr>
<td></td>
<td>happiness, anger).</td>
<td></td>
</tr>
<tr>
<td>SA_5</td>
<td>I express my feelings in appropriate ways.</td>
<td>.06</td>
</tr>
<tr>
<td>Self-management</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SM_1</td>
<td>I use strategies to calm myself down.</td>
<td>.04</td>
</tr>
<tr>
<td>SM_2</td>
<td>I have an appropriate reaction when I'm corrected.</td>
<td>-.06</td>
</tr>
<tr>
<td>SM_3</td>
<td>I control myself while waiting for my turn (no cutting in line, no</td>
<td>.06</td>
</tr>
<tr>
<td></td>
<td>interrupting, etc.).</td>
<td></td>
</tr>
<tr>
<td>Perseverance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PR_4</td>
<td>I stay motivated when doing schoolwork.</td>
<td>-.03</td>
</tr>
</tbody>
</table>

(Table continues)
<table>
<thead>
<tr>
<th>Item ID</th>
<th>Item Text</th>
<th>SA</th>
<th>SM</th>
<th>PR</th>
<th>BB</th>
<th>CF</th>
<th>SF</th>
<th>CFA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PR_5</td>
<td>I keep going, even when the schoolwork is not easy for me.</td>
<td>−.01</td>
<td>−.01</td>
<td>1.00</td>
<td>−.09</td>
<td>.07</td>
<td>.86</td>
<td></td>
</tr>
<tr>
<td>PR_6</td>
<td>I keep working even when life feels hard.</td>
<td>.13</td>
<td>.03</td>
<td>.65</td>
<td>.19</td>
<td>−.01</td>
<td>.00</td>
<td>.92</td>
</tr>
<tr>
<td>BB_1</td>
<td>I bounce back even if I get a bad grade.</td>
<td>.03</td>
<td>.09</td>
<td>.55</td>
<td>.08</td>
<td>.15</td>
<td>−.07</td>
<td>.76</td>
</tr>
<tr>
<td></td>
<td><strong>Bounce-back</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BB_2</td>
<td>Even if I fail at something more than once, I feel OK about myself.</td>
<td>−.04</td>
<td>.06</td>
<td>.11</td>
<td>.74</td>
<td>.08</td>
<td>−.01</td>
<td>.86</td>
</tr>
<tr>
<td>BB_4</td>
<td>I don’t let others’ unkind words make me feel sad for too long.</td>
<td>.08</td>
<td>.00</td>
<td>−.14</td>
<td>.85</td>
<td>−.07</td>
<td>.11</td>
<td>.77</td>
</tr>
<tr>
<td>BB_5</td>
<td>When life is hard, I stay strong.</td>
<td>.01</td>
<td>.09</td>
<td>.15</td>
<td>.66</td>
<td>.09</td>
<td>.01</td>
<td>.92</td>
</tr>
<tr>
<td></td>
<td><strong>Confidence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CF_1</td>
<td>I believe I have talents.</td>
<td>.05</td>
<td>−.02</td>
<td>−.05</td>
<td>−.03</td>
<td>.98</td>
<td>−.02</td>
<td>.85</td>
</tr>
<tr>
<td>CF_2</td>
<td>I know what my strengths are.</td>
<td>.00</td>
<td>.07</td>
<td>.01</td>
<td>.04</td>
<td>.75</td>
<td>.08</td>
<td>.88</td>
</tr>
<tr>
<td>CF_6</td>
<td>I believe I will succeed in life.</td>
<td>.06</td>
<td>.01</td>
<td>.18</td>
<td>.09</td>
<td>.48</td>
<td>.13</td>
<td>.88</td>
</tr>
<tr>
<td></td>
<td><strong>Safety</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF_3</td>
<td>I feel safe at school.</td>
<td>−.03</td>
<td>.09</td>
<td>.06</td>
<td>−.09</td>
<td>.06</td>
<td>.81</td>
<td>.85</td>
</tr>
<tr>
<td>SF_4</td>
<td>I feel safe to approach adults at school with questions or concerns.</td>
<td>−.05</td>
<td>−.01</td>
<td>.03</td>
<td>.08</td>
<td>.02</td>
<td>.76</td>
<td>.80</td>
</tr>
<tr>
<td>SF_5</td>
<td>I feel safe from student online bullying (over the computer, smart phones, social media).</td>
<td>.09</td>
<td>.12</td>
<td>−.02</td>
<td>−.05</td>
<td>−.00</td>
<td>.73</td>
<td>.81</td>
</tr>
<tr>
<td>SF_6</td>
<td>I feel safe to share my thoughts at school.</td>
<td>.04</td>
<td>−.14</td>
<td>.01</td>
<td>.14</td>
<td>.04</td>
<td>.77</td>
<td>.83</td>
</tr>
</tbody>
</table>

*Note.* Bolded factor loadings signify a *p* value <= .05 for an item on its targeted factor. Italics signify non-target loadings with a *p* value <= .05.
Figure 3

Path Diagram Depicting the Structure of the Shortened SEWS Beta Form A
Bounce-back there are three items with factor loadings ranging from .77 to .92. Confidence also consisted of only three items, with factor loadings which ranged from .85 to .88. Finally, we retained four items in the Safety factor with factor loadings that ranged from .80 to .85.

**Measurement Invariance**

Next, I tested for measurement invariance across gender categories (male and female) and grade-level sub-groups (junior high school and high school). As shown in Table 5, CFI dropped from .964 to .96 (< .01) between the configural and scalar models for gender, indicating invariance across the male and female subgroups. In the test for gender invariance, the metric model did not converge. However, because the scalar model is stricter than the metric model, and because the difference between configural and scalar met our a priori thresholds for invariance, I accept this as evidence of measurement invariance. The configural model for grade-level subgroups resulted in a .97 CFI, which dropped to .969 in the scalar model, providing evidence of measurement invariance across school levels.

**Reliability**

Finally, I also estimated the reliability of each subscale. Because the assumption of tau-equivalence was not met, I used the McDonald’s model-based omega coefficient. Table 6 shows that most subscales had a high reliability coefficient (> .88). Only Self-management had a reliability less than .80. This is evidence that the results collected are internally consistent at a subscale level.
Table 5

*Invariance for Gender and School Level CFA*

<table>
<thead>
<tr>
<th>Invariance Test</th>
<th>CFI</th>
<th>TLI</th>
<th>RMSEA</th>
<th>SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Configural Model</td>
<td>.964</td>
<td>.958</td>
<td>.063</td>
<td>.045</td>
</tr>
<tr>
<td>Metric Model</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Scalar Model</td>
<td>.960</td>
<td>.960</td>
<td>.062</td>
<td>.048</td>
</tr>
<tr>
<td><strong>School</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Configural Model</td>
<td>.970</td>
<td>.965</td>
<td>.065</td>
<td>.045</td>
</tr>
<tr>
<td>Metric Model</td>
<td>.970</td>
<td>.966</td>
<td>.063</td>
<td>.045</td>
</tr>
<tr>
<td>Scalar Model</td>
<td>.969</td>
<td>.969</td>
<td>.060</td>
<td>.047</td>
</tr>
</tbody>
</table>

*Note.* While the metric model for gender invariance did not converge, the scalar model converged with less than a .01 reduction in CFI between the configural and scalar models.

Table 6

*Reliability Estimates by Subscale*

<table>
<thead>
<tr>
<th>Subscale</th>
<th>McDonald's Omega</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perseverance</td>
<td>.90</td>
</tr>
<tr>
<td>Confidence</td>
<td>.90</td>
</tr>
<tr>
<td>Self-awareness</td>
<td>.90</td>
</tr>
<tr>
<td>Safety</td>
<td>.89</td>
</tr>
<tr>
<td>Bounce-back</td>
<td>.88</td>
</tr>
<tr>
<td>Self-management</td>
<td>.78</td>
</tr>
</tbody>
</table>
Previous research has suggested that social and emotional learning (SEL) can have a meaningful impact on academic achievement as well as personal development (Durlak et al. 2011). Implementing school wide SEL curricula and measuring their impact is important but challenging. By collaborating with researchers from Brigham Young University, Alpine School District developed the Social and Emotional Well-being Scale (SEWS), a self-report scale intended to assist teachers and principals in the process of choosing appropriate SEL interventions and accurately evaluating the effects of the chosen intervention at a school-wide level. The purposes of the current study were to (a) assess the psychometric properties of the SEWS Beta Form A in terms of model fit, reliability, and measurement invariance, (b) reduce the total number of items to create a shorter version, and to (c) investigate the usefulness of comparing ESEM with target rotations and CFA results as part of the scale development process. Based on the iterative factor analysis results, I suggested a modification to the original model. With this modification, the shorter version resulted in acceptable model fit, measurement invariance across gender and school level, and high levels of sub-scale reliability based on the sample obtained.

**Constructs and Internal Structure**

This pilot study found that the proposed model which consisted of six first-order factors included (a) Safety, (b) Confidence, (c) Bounce-back, (d) Perseverance, (e) Self-management, and (f) Self-awareness, and two second-order factors (Resilience and Self-mastery) did not fit the data in a way that resulted in distinct factors. Specifically, Confidence, Bounce-back and Perseverance were so highly related that our initial attempt to shorten the survey
resulted in a lack of discriminant validity between them. Work by Stajkovic et al. (2015) supports these findings. Their model of a general factor of confidence was composed in part of sub constructs like bounce-back and perseverance.

Overall, most of the factors on the shortened version of the SEWS Beta Form A retained the essence of their theoretical definition. One danger in scale development is that in the item elimination process, the goal of retaining items that are empirically consistent with each other can result in a subscale that underrepresents the more qualitative breadth of the construct (Bandalos, 2018). In the following sections, I will review the definitions proposed in the literature review and used in item generation for each construct and evaluate to what degree the surviving items measure the breadth and depth of the construct.

**Self-Awareness**

The construct of Self-awareness was defined as the degree to which students were aware of their emotions and related to the construct of emotional understanding in the literature. MacCann et al. (2020) defined emotional understanding as the knowledge of words used to describe emotions, and the understanding of their causes and consequences. The retained items conform to this definition and each item contributes unique information. For example, item SA_1 (e.g., “I notice when my feelings change (happy to sad, angry to calm).”) differs from SA_5 (e.g., “I notice what my body does when I experience emotions”) in that SA_1 focuses on noticing the moment of change, while SA_5 focuses on the effects of the change. Other items cover why an emotional change has occurred, and the ability to articulate an emotional condition. Combined, the items fairly represent important aspects of emotional self-awareness or emotional understanding.
**Self-Management**

The construct of Self-management was theorized to be the degree to which a student could regulate their emotions and was patterned after the construct of emotional management form the emotional intelligence literature. In the literature, emotional management was described as the ability to regulate and respond strategically to emotions (MacCann et al., 2020). The content of the four items that tap into the degree to which students (a) express themselves appropriately, (b) use strategies to remain calm, (c) respond well to correction, and (d) wait patiently for their turn. These four scenarios highlight observable behaviors and unique contexts in which students may express emotional self-control.

**Perseverance**

The construct of Perseverance was defined as the ability to endure difficulty over a period of time. Specifically, items were written with endurance beyond a single class period in mind, but shorter than the much longer-term perseverance described by Duckworth et al. (2007). The four items that were retained invite students to rate their ability to (a) stay motivated during the school day, (b) endure difficult schoolwork, and (c) bounce-back after a bad grade (BB_1). The underlying themes in these items are (a) endurance, (b) resilience, and (c) action, particularly in the context of schoolwork. After omitting items that cross-loaded strongly between Bounce-back and Perseverance, we discovered that there was an emotional component to the retained Bounce-back items that did not exist in the perseverance items. However, BB_1, which modification indices had suggested belonged to the Perseverance factor was missing this emotional aspect (“I bounce-back even if I get a bad grade.”). In other words, even though the term Bounce-back appears in a Perseverance item, it relates well with the type of resilience that the other items of Perseverance define.
**Bounce-Back**

The construct of bounce-back had been associated with academic buoyancy, or the ability to cope with everyday challenges (Martin & Marsh, 2009). However, an examination of the content of the remaining items in the Bounce-back factor revealed a different underlying theme. Only three items for Bounce-back differentiated themselves from the constructs of Perseverance and confidence. These three items tap heavily into the student’s ability to respond with emotional stability in the face of (a) failure, (b) unkind words, or (c) general difficulty. However, while this emotional theme was not originally a key part of our definition, or the definition of academic buoyancy, it made sense as an important and distinct social and emotional skill to measure.

**Confidence**

The construct of confidence was defined as the degree to which a student believed in themselves and their abilities to succeed. I related this construct generally to a variety of constructs related to positive thinking (Anderson et al., 2016), but specifically to the core confidence construct recently investigated by Stajkovic et al. (2015). Core confidence subsumed optimism, hope, and bounce-back, and is manifested by an assurance of what to do, a belief in one’s ability to do it, and a resilience to failure. Specifically, Stajkovic et al. (2015) proposed a hierarchical model, with Core-confidence as a second-order factor.

The three items that were retained measure a student’s belief that they have talents, their belief that they will succeed in life, and their ability to identify those talents. While the literature review of this construct led us to suspect that the ASD definition of confidence could incorporate elements of optimism, general self-efficacy, and hope, the items that were retained only reflect optimism (e.g., “I believe I will succeed in life”) and general self-efficacy (e.g., “I believe I have talents,” and “I know what my strengths are”). The action-oriented and goal-oriented qualities of
hope were too highly related with perseverance. For example, “I am confident I can perform as well as other students in school” and “I believe I can accomplish anything I set my mind to” both cross-loaded onto the perseverance construct (.20, .18) in the first iteration of the model. These loadings were small but statistically significant and may hint at the relationship between hope and perseverance. Additionally, Confidence was originally intended to be a construct of condition, like Safety. Further review of the retained items indicated it is more like other skill constructs (e.g., Self-management, Self-awareness, etc.) which measure an ability or trait.

**Safety**

While there have been several approaches to measuring school safety, we attempted to measure student perception. Specifically, the construct of Safety was defined as the degree to which students felt physically and emotionally safe at school. However, none of the four items that were retained specifically touched on the physical aspect of safety. Rather the content of the items addressed (a) feeling safe at school, (b) feeling safe to get help from adults, (c) feeling safe to share thoughts, and (d) feeling safe from online bullying. We found these to be sufficiently broad while at the same time tapping into the conditions of safety in a school context.

**Psychometric Qualities**

Beyond the content of the items and subscales, there were several psychometric properties which provide evidence of the instrument’s potential usefulness. The overall fit of the model was acceptable, and the ESEM and CFA loadings provided strong evidence that the remaining items loaded well onto their designated factors. Additionally, the correlations among factors were lower than .80, especially in the ESEM model which is less likely to include inflated correlations. Together these pieces of evidence support the use of these scores as estimates of the six constructs from the model. The high reliability coefficients also provided some assurance of
consistent scoring, meaning that given a different but similar sample of items, students would likely have responded in similar ways. Scores obtained from a scale are meant to be a snapshot. The reliability of this snapshot determines its usefulness in generalizing beyond these items. Finally, the evidence of measurement invariance across gender and school level allows differences in scores at these group levels to be interpreted as actual group differences, providing important insight into the student population. Evidence of measurement invariance is also important as an assurance that the scores are explained by the same factor structure, or that the raw scores and latent scores are not systematically different by group.

Together these disparate pieces of evidence support two main uses and interpretations of the scores. The internal structure suggests that items associated with each construct can be aggregated at the subscale level (e.g., items loading on Safety can be averaged or summed), and that the composite score will be a relatively accurate representation of a student’s trait level for that construct. Secondly, the mean differences between males and females, as well as the mean differences between school levels can be compared and interpreted as substantive differences in the population. In contrast, there are several limitations on how these scores can be used or interpreted. There is insufficient evidence for the aggregation of first-order construct scores into a composite of second-order construct scores like Resilience or Self-mastery. Additionally, the aggregation of sub-scale scores, or comparison of group mean differences would be limited to those obtained in this pilot study, as the participating students did not sufficiently represent the diversity of ASD’s student population.

**Exploratory Structural Equation Modeling and Target Rotation**

Another objective of this research was to investigate the usefulness of comparing ESEM and CFA results, specifically when target rotation was applied. ESEM with target rotation allows
a theoretical model to be imposed on an Exploratory Factor Analysis, like the theory driven approach of CFA. However, rather than constraining all non-target loadings to be zero, the target rotation approaches zero on non-target loadings, while still allowing non-zero results. As has been shown in other studies comparing results from ESEM and CFA models, the correlations among factors and target loadings on target factors tended to be smaller in the ESEM (e.g., Booth & Hughes, 2014; Marsh et al., 2010; Gomes & Gjikuria, 2017). Additionally, the ESEM had more acceptable fit statistics. I did not notice any useful differences in the modification indices between ESEM and CFA. On some iterations, the modification indices for the ESEM did not seem to have been estimated properly (.999 M.I. for by statements between all items, but the E.P.C was 0).

Both the CFA and ESEM informed the iterative process of omitting items to make the shorter version of the SEWS Beta Form A. The CFA results helped us identify items that achieved high loadings on the target factor, while the ESEM provided insight into salient cross-loadings. The salient non-target cross-loadings provided us with a sense of how much an item was contributing to overall model misfit even if in the CFA it loaded highly on its intended factor. This information was supplemented by but distinct from the modification indices, which some scholars have warned against using mechanically in small samples, as such a search capitalizes on chance (MacCallum, 1992). Comparing the correlations among factors in the final ESEM and CFA models also provided us evidence of the degree to which constraining cross loading items to 0 was inflating the correlations between factors. In this case, only marginal inflation occurred, which supports the use of the more parsimonious CFA model.

Some have suggested that CFA is too strict for complex and closely connected psychological and social constructs, and that ESEM provides a more reasonable factor analysis
(e.g., Marsh et al., 2014; Booth & Hughes, 2014; Perry et al., 2015). In our case, the pilot study data fit the CFA model, which is a more parsimonious version of the ESEM. However, this may not be the case in other situations. I believe that the real advantage of comparing CFA and ESEM with target rotation is that the researcher can get a sense of the complexity of the data based on which model best explains it. In other words, ESEM may be a more appropriate model in other situations.

For example, Figure 4 provides a visual summary of the proportion of item variance explained by each of the 23 items respectively. The ESEM loadings were squared and then stacked in the corresponding item columns. Each factor contributes to the variance of each item to some degree, as shown by the different colored segments of each column. In most cases, a significant portion of each item is influenced by a single factor, and the influence of other factors are minimal. Figure 4 provides an intuitive sense for the relative size of what I consider small cross-loadings.

However, Marsh et al. (2014) and proponents of more flexible modeling argue that allowing these small or non-significant cross-loadings to occur, results in parameter estimates, models, and scores from instruments that are more replicable across populations (Morin et al., 2016; Asparouhov & Muthén, 2009). In some senses, items and models that reflect the complex landscape of psychological and educational constructs are bound to have multidimensional components. Morin et al. (2016) argued that these components would even be construct-relevant, meaning the attempt to model them in as pure indicators of a factor would undermine their validity. Modeling expected complexity is less convenient, but perhaps more accurate and ultimately more useful. As McCrae et al. (1996) argued, there is little scientific utility in an instrument or model that cannot be replicated.
As an indicator of how allowing small non-target cross-loadings to occur influences our understanding of each items’ relationships with the model factors, and the degree to which they contribute meaningful information, consider CF_1 and CF_6. Almost 100% of CF_1 (i.e., “I believe I have talents”) can be explained by the confidence factor, while only about 25% of the variance of item CF_6 (i.e., “I believe I will succeed in life.”) is influenced by confidence. This is in stark contrast to the variance explained in the CFA model, where CF_6 had a loading of .87 on confidence, resulting in 75% of the variance being explained by confidence. The ESEM loadings provide important insight into the relationship between items and their intended factors in the presence of the other factors, rather than the items and factors siloed.

Because of the similarities between ESEM with target rotation and a traditional EFA with an oblimin rotation like goemin, I also compared factor loadings, correlations among factors, and
fit statistics between the two for the final model. The EFA returned a CFI of .99, a TLI of .98, an RMSEA of .05, and an SRMR of .02. The correlations among factors tended to be systematically greater in the ESEM with target rotation ($M = .04, SD = .02$) than factor loadings in the EFA with geomin rotation. The factor loadings were also systematically greater in the ESEM with target rotation ($M = .005, SD = .02$). However, the degree of difference between the two estimates was so small that I considered them to be essentially equivalent. A simulation study conducted by Xiao et al. (2019) indicated that target rotation more accurately estimated the main loadings, and correlations among factors, while it did worse than geomin in estimating cross-loadings. Based on Xiao et al.’s findings, I interpret the lack of differences to suggest that our theoretical findings are well supported by the more data driven rotation methods.

**Limitations**

One of the primary limitations of this study was the limited sample size. While I was able to obtain a sample that exceeded the minimum number of recommended participants for a CFA model (i.e., 5 students per item, and more than 100 observations), the sample was still relatively small for so many parameter estimates (Brown, 2015). It also was ethnically homogenous, and only represented participants from a certain geographical area in the school district. These limitations placed constraints on our ability to generalize these results to all ethnicities and all secondary level schools in ASD. Closely related to this issue, the small amount of missing data was highly concentrated in the junior high level. While it is unclear the amount of bias this introduced to our estimates, I acknowledge this limitation. Finally, the limited sample size prevented me from splitting the sample in half to cross-validate the results. Because I used an exploratory approach to reduce the number of items in the scale, the current model is only a
tentative model. It is possible that this model has been tuned to the sample and will not function well on a different set of data.

Additionally, there are several limitations on the kinds of valid interpretations and uses of current scores. One of the intended purposes of the scale is to measure progress. This study did not attempt to test for invariance over time or to assess test-retest reliability. I also suggest a more thorough investigation into the malleability and meaningfulness of each construct as they relate to additional criteria. In other words, to what degree, if any, do these constructs respond to intervention and to what degree, if any, do they predict beneficial outcomes on other variables (e.g., academic, or social outcomes). For example, outcomes such as (a) homework completion, (b) standardized testing scores, (c) grade-point-average, (d) suspension, (e) attendance, (f) extracurricular participation could all be incorporated into an ESEM or SEM model, which could give insight into the causal or mediating influence of the SEWS Beta Form A constructs on student well-being and success. This would in turn provide guidance for teachers and principals about the constructs of greatest impact both within ASD and potentially outside of ASD. It would also provide insight into the degree to which the SEWS Beta Form A scales relate to school outcomes in ways that parallel or diverge from other scales in the literature. However, until these studies are conducted, scores can only be interpreted as a snapshot of current students’ self-perceptions.

Another possible use of the scale in the future could be as a tier-one screener to help identify internalizing students who may need additional support. Before using the results of this scale for that purpose, it would be important to implement some version of known groups testing. In this case, ASD could identify one group of internalizing students that are receiving additional services from a school counselor or school psychologist and compare their results with
a second group of students who do not need additional support. If SEWS scores can accurately predict or identify which student belongs to which group, then the scores could be used to help teachers and principals identify students that need help but who may go unnoticed.

While the original model and item pool were generated based on a review of the literature and the practical expertise of the ASD school psychologists, this was ultimately an exploratory project. A true confirmatory approach should be conducted, preferably with a larger sample that better represents the district population. Additionally, reliability by group at the subscale level should be estimated.

Also, the validity of interpreting these scores as being reflective of their target constructs could further be supported by additional stakeholder and expert input. For example, the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014) recommends consulting experts in relevant fields to help evaluate the validity of the content. This could include teachers, parents, and researchers assessing the degree to which the items sufficiently represent the construct or identifying items that introduce content-irrelevant variance (e.g., items with language that taps into cultural knowledge, items that measure skills or behaviors irrelevant to the construct). Consulting a variety of stakeholders and experts fosters consensus on the meaning of items. To my knowledge, ASD plans on holding open forums to allow parents the opportunity to provide feedback on the items as they move forward with the scale.

**Recommendations**

I recommend the use of ESEM with target rotation in conjunction with CFA during the scale development process for measures of SEL. ESEM has already been used in a variety of similar fields of research such as (a) personality research (Boffo et al., 2012; Marsh, Nagengast,
& Morin, 2013; Neff et al., 2019), (b) child behavior research (Hukkelberg et al., 2018), (c) mental health research (Joshanloo, 2016b, 2018) and (d) exercise science research (Garn & Webster, 2018; Hoffmann & Loughead, 2019). Like in these fields of study, SEL scales tend to include constructs that are highly correlated, making ESEM with target rotation an ideal exploratory approach to developing appropriate instruments. Researchers and practitioners will often already have a theoretical model in mind. ESEM allows the initial stages of development to incorporate this theory while still providing informative estimates of the degree to which the theory is violated at an item level. While the more constrained CFA approach provides a simpler and idealistic model, a sound argument can be made for the use of ESEM as the primary indicator of the validity of the internal structure. In this study, cross-loadings were small, and most of them were not statistically significant. The ESEM model provides a more realistic description of the interrelatedness of the SEL items and constructs in the SEWS Beta Form A, because common variance between items intended to measure different constructs is accounted for at the correct level (i.e., the item level), rather than at the factor level.

I also recommend the use of ESEM when assessing the relationships between constructs and other criteria. As described previously, shedding light on mediating, moderating, or causal effects of SEL on academic and other well-being outcomes empowers practitioners to make informed decisions about which constructs matter and to what degree. If constructs are closely related, for example they share a correlation between factors of .80 or close to it, then using those constructs as predictors or mediators of GPA will potentially result in inflated standard errors due to effects of multicollinearity. As shown in our study, the ESEM model returns lower correlations among factors. In other words, as has been suggested in other research, extending the flexibility of an EFA into the realm of an SEM may result in more accurate understandings of
the relationships between the measurement component and the structural component (Marsh et al., 2010; Marsh, Nagengast, & Morin, 2013; Joshanloo, 2016b)

At the very least, the comparison of the ESEM model with the CFA model gives insight about the degree to which correlations among factors are inflated in the CFA model, a known issue that is particularly troublesome in the development of scales with closely related constructs. It is possible that some SEL measures are discarded because of a lack of evidence of discriminate validity between constructs, when the real issue lies at the item level in the form of small and insignificant cross-loadings.

However, while I argue that when cross-loadings are small, ESEM with target rotation provides valuable and viable parameter estimates, I agree with previous researchers that the most parsimonious model should be used (Marsh, Nagengast, & Morin, 2013; Marsh et al., 2014; Asparouhov & Muthén, 2009). Perhaps better put, the simplest model that best reflects theory in a way that is useful should be used. Sometimes the data is better explained by a simple model, and sometimes it is better explained by a more complex model.

A follow up study should be performed to determine the impact of the collaborative approach to this research project. Our hypothesis is that partnering with educational practitioners will result in more consistent progress in educational research and outcomes. As researchers and practitioners find intersections of interest, they are more likely to collaborate, experiment, iterate, retain insights, and pass on these insights to newcomers. I hope that ASD will continue to partner with researchers either from BYU or other universities to continue the development of the SEWS Beta Form A and B and increase capacity within their district to effectively use the data they obtain. I also encourage this relationship because it clears the way for larger scale research into the relationships and definitions of SEL constructs that will benefit all parties involved.
Currently, ESEM is limited in several ways. As Marsh et al. (2014) points out, ESEM allows researchers to integrate an EFA into MIMIC models, longitudinal studies, and test for measurement invariance, but it does not readily facilitate the integration of second-order factors. Software that accomplishes this would expand the usefulness of ESEM, especially in research on SEL which most frameworks assume to consist of tiers of constructs. Also, it is unclear how to estimate reliability coefficients because model-based reliability coefficients like McDonald’s omega are made for the simpler models that result from CFA. Further research into how reliability measures can be adapted to meet this type of model would be valuable.

Conclusions

The SEWS Beta Form A scale was developed in a collaborative effort between ASD and researchers at BYU. The self-report scale was designed to measure (a) Safety, (b) Confidence, (c) Bounce-back, (d) Perseverance, (e) Self-management, and (f) Self-awareness as first order constructs, and (a) Resilience and (b) Self-mastery as second order constructs. After reviewing the literature, generating a pool of items, investigating construct validity, and revising items, a scale of 38 items was administered to students at the junior high school and high school levels. Using ESEM with target rotation in conjunction with CFA, we iteratively omitted items and conducted factor analyses to create a shorter version of the SEWS Beta Form A scale. The final version consisted of 23 items and an adjusted model, removing the second-order factors. In both the CFA and ESEM models, global fit statistics were acceptable. Factor loadings were high in the CFA model, and the ESEM model achieved low secondary loadings, and high factor loadings on primary factors. I obtained evidence of measurement invariance across gender and school level, and high reliability index scores at the subconstruct level. Together, this preliminary evidence supports the use of these scores as indicators of students’ self-perceived ability levels.
for each construct measured. Evidence also supports the use of the scores to compare group differences in terms of grade level or across gender subgroups. Further research should be conducted to confirm the results of this pilot study with a larger and more representative sample. Comparing ESEM and CFA proved to be provided needed insight that facilitated the development of this instrument. Finally, I highly recommend that scales be developed in collaboration with practitioners and stakeholders to provide needed insight and feedback into its structure and use.
References


https://doi.org/10.1207/s15326985ep4104_4


https://doi.org/10.1177/1094428114555994


https://doi.org/10.1080/1091367X.2015.1081594


https://doi.org/10.1016/j.jrp.2007.02.001


Day, L., Hanson, K., Maltby, J., Proctor, C., & Wood, A. (2010). Hope uniquely predicts objective academic achievement above intelligence, personality, and previous academic
https://doi.org/10.1016/j.jrp.2010.05.009


https://doi.org/10.1111/j.1467-8624.2010.01564.x


https://doi.org/10.5334/pb.369


Muenks, K., Wigfield, A., Yang, J. S., & O’Neal, C. R. (2017). How true is grit? Assessing its relations to high school and college students’ personality characteristics, self-regulation,

https://doi.org/10.1037/edu0000153


https://bit.ly/3tlkHd1


https://files.eric.ed.gov/fulltext/ED364652.pdf#page=132


https://doi.org/10.1016/j.appdev.2017.06.001


https://doi.org/10.1186/1477-7525-9-8

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach’s α, Revelle’s β, and McDonald’s ωH: Their relations with each other and two alternative conceptualizations of reliability. Psychometrika, 70(1), 123–133. https://doi.org/10.1007/s11336-003-0974-7
APPENDIX

Consent/Institutional Review Board Approval Letter

Memorandum

To: Richard Sudweeks
Department: BYU - EDUC - Instructional Psychology & Technology
From: Sandee Alva, MPA, HRPP Associate Director
Wayne Larsen, MAsc, IRB Administrator
Date: October 01, 2021
IRB#: IRB2021-264
Title: The Development of a Social and Emotional Well-being Scale Using ESEM and CFA: Synergistic Stories in Complex Models

Brigham Young University’s IRB has approved the research study referenced in the subject heading as exempt level, category 4. This study does not require an annual continuing review. Each year near the anniversary of the approval date, you will receive an email reminding you of your obligations as a researcher and to check on the status of the study. You will receive this email each year until you close the study.

The study is approved as of 10/01/2021. Please reference your assigned IRB identification number in any correspondence with the IRB.

Continued approval is conditional upon your compliance with the following requirements:

1. A copy of the approved informed consent statement can be found in IRIS. No other consent statement should be used. Each research subject must be provided with a copy or a way to access the consent statement.
2. Any modifications to the approved protocol must be submitted, reviewed, and approved by the IRB before modifications are incorporated in the study.
3. All recruiting tools must be submitted and approved by the IRB prior to use.
4. Instructions to access approved documents, submit modifications, report adverse events, and/or informed consent processes. Such modifications require the review and approval of the IRB. Please refer to the IRB website for more information.
10/13/2021

Christopher Busath

Dear Christopher Busath,

Thank you for your interest in conducting your research with Alpine School District. I am granting you permission in conjunction with Melissa Bostwick, Kim Jones and Vallen Thomas for the project you requested.

Good luck in your research, and if you have any questions, please don’t hesitate to call at the number stated above.

Sincerely,

Gary Twitchell

GT/TS

Director of Research and Evaluation