Using psychometric models to measure social and emotional learning constructs

A Dissertation
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Mireya Carmen-Martinez Smith

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOHY

Michael Rodriguez, Adviser

August 2020

Acknowledgements

I am forever grateful to my adviser, Dr. Michael Rodriguez, who welcomed me in the measurement community and taught me how to think as a methodologist/ psychometrician. Thanks to Dr. Rodriguez, I finally found a safe place to learn – in the Quantitative Methods in Education (QME) department where Dr. Robert delMas, Dr. Ernest Davenport, and Dr. Andrew Zieffler were generous with their time in explaining quantitative concepts and shared very useful books/articles that expanded my knowledge in QME. Furthermore, I am grateful for having the privilege of working as a research assistant at the Research Methodology Consulting Center (RMCC) with extremely supportive colleagues and an excellent supervisor, Dr. Danielle Dupuis who valued my quantitative work and showed me how to be a QME woman researcher. My time in QME program was more enjoyable thanks to being part of the Minnesota Youth Development Research Group led by Dr. Rodriguez where I learned from other graduate students and collaborated on papers to present at conferences. To my lifelong QME friends, I hope we continue sharing the joys and challenges of QME research. Thank-

you, Dr. Vichet Chhuon for sharing your expertise in racial and ethnic identities.

This journey started with my mom (Florinda Carlos) who taught me how to read; my dad (Victor Martinez) who showed me the joy of learning and reading; my sister (Priscilla Martinez-Carlos) who taught me how to be brave; and my aunt (Dra. Elvia Carlos) who showed me that women can be doctors too! Thanks to my close friends who cheered me on throughout my 3.5 years in the PhD program: Krista, Lidan, Marian, Crystal and Khanh Keilo (my 9-yr-old son), who gave me the strength and motivation to finish this degree – now we can enjoy more hours of reading/playing together.

Dedication

To my best-friend, partner, writing editor, and technology support: Garett Smith, who saw it was worth making all the sacrifices for me – so that I could live out my dream of pursuing and completing a doctorate degree.

Abstract

In *Testing Standards* (2014), a construct is a concept or characteristic that an assessment is intended to measure. From a quantitative lens, a construct is trait or domain that may include attitudes, skills, abilities, dispositions and some aspects of knowledge (e.g., competencies). Research studies suggest that social and emotional learning (SEL) constructs may be useful in narrowing the achievement gap, however there is no agreed upon definition of SEL as SEL constructs are multifaceted and defined by the researcher(s). Currently, some SEL constructs are measured qualitatively but this ignores the quantitative structure of the construct. In the quantitative field, SEL constructs are measured by applying a less complex model before a complex model. However, this disregards the qualitative definition for the SEL construct. Furthermore, a construct cannot be directly measured (e.g., person's height), instead, we need to indirectly observe SEL constructs through item responses (e.g., polytomous items).

The problem is that there is a lack of clarity in how the SEL constructs are defined and measured. In addition, there is very little research in an approach for SEL constructs to have accumulating evidence that supports score interpretation and use. This study proposes using the paradigm for SEL assessment that can lead to meaningful, useful, appropriate, and fair score interpretation and use. The paradigm consists of three components. The first component, the structural components of SEL, makes a distinction of the units of SEL assessment (framework, construct(s), measure(s) and item responses) where the construct is the centerpiece. The second component is where the construct definition and measurement model work together to put forth plausible competing models

for the internal structure (e.g., bifactor) of selected SEL constructs. The final component is forms of validity evidence (e.g., measurement invariance) where the focus is to evaluate the claims (e.g., scores can be compared across groups) regarding what the scores represent and how they should be used. The paradigm for SEL assessment encourages researchers from the qualitative and quantitative fields to work together to properly define SEL constructs in a qualitative (e.g., theory) and quantitative (e.g., confirmatory factor analysis and item response theory models) manner.

Table of Contents

List of Tables

List of Figures

Chapter I: Introduction

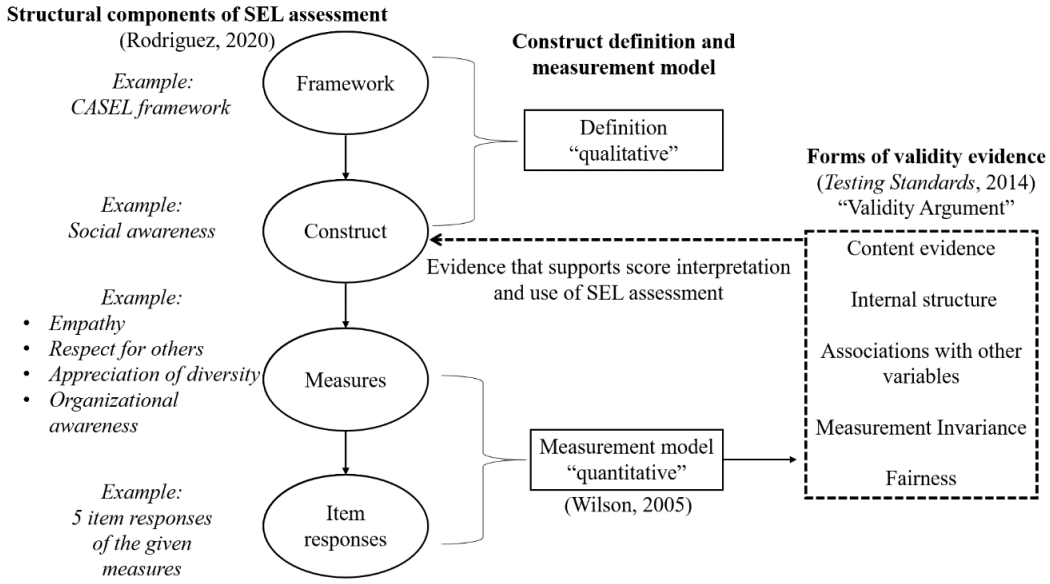
Social and emotional learning (SEL) constructs have received a lot of attention in research, practice, and policy evidenced through the vast amount of available SEL resources and published work of large-scale projects (Berg et al., 2017; Berman, Chaffee, Sarmiento, 2018; Cross-agency project team, 2018; Domitrovich, Durlak, Staley, Weissberg, 2017; Gabrieli, Ansel, Krachman, 2015; Gehlbach & Hough, 2018; Hamilton, Stecher, Schweig, Baker, 2018; Jones, 2018; Restorative Practices Working Group, 2014; Taylor et al., 2018; Yoder, 2014). SEL competencies are believed to be malleable (Farrington et al., 2012), have a positive effect on academic outcomes (Durlak, Weissberg, Dymnicki, Taylor, Schellinger, 2011), predict future academic performance (Ross & Tolan, 2018; Panayitou, Humphrey, & Wigelsworth, 2019) and SEL skills are positively associated with important life outcomes (Chernyshenko, Kankaras, & Drasgo, 2018)**.** There is no one agreed upon definition for SEL as the SEL constructs are multi-faceted and may include cognitive, social, & emotional skills and competencies (Jones & Khan, 2017). Thus, there needs to be more clarity in how the SEL constructs are defined and measured (Buros-Spencer Project Scholars, 2020). Furthermore, the SEL constructs need to have accumulating evidence that supports score interpretation and use (Rodriguez, 2018).

In psychology, a multifaceted construct is a psychological attribute/trait of a person that cannot be directly measured like a person's weight (Crocker & Algina, 1986). For decades, researchers/psychometricians/psychologist have offered different approaches to measuring multifaceted constructs in personality psychology (Hull,

Lehn, & Tedlie, 1991). Take for example a person's trait, an attitude towards a topic, a construct, that has been measured by Likert (1932) and Thurstone and Chave (1929) using a total score approach (e.g., summing the item score with equal and unequal weights). Similarly, in education, a multifaceted construct that cannot be directly observed, such as academic achievement, must rely on items/tasks to indirectly observe the construct. In general terms, a construct is operationalized through measures which are composed of items (e.g., survey) that are relevant to the construct. A construct may be part of an assessment.

The structural component of SEL assessment consists of four different levels (from the largest unit to the most specific in units): Framework, construct, measures and item responses. Each level has its own definition and are distinct from each other (Rodriguez, personal communication, April 14, 2020). The framework is the qualitative structure of how a construct is conceptualized and operationalized whereas the construct is the more specifically defined latent trait (may include skills, abilities, dispositions and some parts of knowledge, also known as competencies and domains). The construct is quantitatively operationalized by a specific tool – a measure or measures. These measures are comprised of items or tasks that are a way to indirectly observe a person's position on the construct being measured. The framework, construct, measures and item responses are depicted in Figure 1.1 as ovals.

Figure 1.1
Paradigm for SEL Assessment



To give an example, the Collaborative for Academic, Social, and Emotional Learning (CASEL) is the framework (the qualitative structure), which describes how a construct such as Social Awareness is proposed; where Social Awareness may be quantitatively defined by four measures (*Empathy, Respect for others, Appreciation of diversity, & Organizational awareness*) comprised of five items from a survey, that result in item responses—the observed manifestation of the construct, as situated in the SEL framework. The better the construct is qualitatively defined in a plausible framework, the better the construct can be quantitatively measured, ultimately leading to valid (meaningful, useful, appropriate, and fair) interpretation and use.

The framework formulates the qualitative definition of the construct, proposing how one or more constructs are organized and associated, as a part of a nomological net. The definition of a construct is theory driven. The other two parts of the SEL assessment paradigm, measures and item responses, compose the quantitative definition of the model through the measurement model. The definition and measurement model of the construct are depicted as rectangles in Figure 1.1.

The measurement model includes statistical models that use item responses to analyze the scored responses and relate scores back to the construct (Wilson, 2005). One approach (generally consistent with Wilson, 2005) is to define a construct rather simply, such that it can be represented on a continuum from low or little of the trait to high or much of the trait. Then to measure the SEL construct we use psychometric models (e.g., Item Response Theory) where the SEL measure(s) is/are created by aggregating rating-scale items (e.g., *strongly disagree*, *disagree*, *agree*, *strongly agree*). The assumption is the participant possesses some amount of the construct that causes the responses of the items. We score the item responses (e.g., in an outcome space such as 0 = false and 1 = true) and use a psychometric model (e.g., Rasch model) to infer about the person's position on the underlying construct. Note that the psychometric models need to be plausible models that agree with the qualitative structure of the construct which is important for score interpretation and use.

Another psychometric model to measure SEL constructs is confirmatory factor analysis. The factors are also known as domains which are constructs. Usually, the percent of a domain is reported to represent how much of a domain the

participant/student has mastered (Rodriguez, personal communication, May 10, 2020). Item response theory (IRT) and confirmatory factor analysis (CFA) are two theoretical approaches to measure constructs which Reckase (2017) argues result in different interpretation of the measurement of the construct. As stated previously, in IRT the construct is measured on the amount of a trait the person displays and in CFA the construct is a measurement of how much the person has obtained. These two different views of measuring a construct may be conflicting where the latent traits approach estimates a person's location on a continuum and the domain approach estimates the proportion of a domain a person has acquired (Reckase, 2017). Reckase (2017) acknowledges one of challenges for the latent traits approach is to use a well-defined continuum and in the domain approach is to obtain a representative sample of the domain. The measurement model includes these two different psychometric models to measure SEL constructs using item responses that compose the measures. As Messick (1989) noted, validity evidence is collected at the measure level and applied to the construct and not to the framework; however, the framework shapes the plausible internal structure of the measure for the construct.

Validity Evidence

The last part of Figure 1.1 contains the forms of validity evidence (dotted rectangle) that directly connect to the construct (dotted arrow). The forms of validity evidence are collected at the measure level (not framework or construct level) and support the score interpretation and use argument (IUA) as proposed by Kane (2013b). The IUA "includes all of the claims based on the scores" and the IUA need to be

"coherent and complete" (Kane, 2013, p. 2,). Some IUAs may have several possible uses while other IUAs may have one particular use. The more the IUA claims, the more forms of validity evidence required. Validity is a unitary concept that supports score interpretation and use (*Testing Standards*; AERA, APA, & NCME, 2014). The goal in this context is to have a collection of evidence to support the interpretation and use of scores for SEL assessments (Rodriguez, 2018).

As "validity is a matter of degree" (p.3, Kane 2013), the process of validation is to accumulate evidence to support the IUA. The many forms of validity evidence particularly useful in the context of SEL may include content evidence, evidence based on internal structure, evidence based on associations with other variables, measurement invariance, and fairness. Here, the five forms of validity evidence are documented as the evidence needed to evaluate the IUA for an SEL assessment.

These five forms of validity are collected at the measure level and the extent of validity evidence supports the IUA at the construct level (Messick, 1989). Content evidence can include an assessment blueprint of how the items map to the SEL construct. The blueprint can be reviewed by a diverse panel of experts, so the content can be used "to address questions about differences in the meaning or interpretation of scores across relevant subgroups" (*Testing Standards,* 2014). Evidence based on internal structure includes the extent to how item loadings contribute to the structure of the SEL construct (Rodriguez, 2018b). The third form of validity evidence is associations to other variables and pertains to the SEL measure indicating strong and positive associations with similar variables intended to measure similar constructs

(convergent evidence) as well as associations between SEL scores from different internal structures (discriminant evidence), and how accurately the SEL construct predicts a criterion measure such as school academic grades (test-criterion associations).

Evidence pertaining to comparing groups, specifically across community groups with different characteristics, includes measurement invariance (MI), and fairness (Rodriguez, 2018). Evidence of MI is derived from the evidence based on response processes stated in the *Testing Standards* (2014). There is also a series of assumptions underlying score interpretation that may employ evidence of scoring, inferences made from item responses to measures, and then to constructs, and generalization inference (e.g., reliability). MI involves the score interpretations to have consistent meanings across groups, especially if groups are being compared. The observed difference in scores must be due to the measure of the construct and not irrelevant to the construct. Fairness is in the realm of evidence based on the intended and unintended consequence of SEL score interpretation and use. It also pertains to having the SEL assessment be inclusive to all students, including students from diverse backgrounds.

The framework has an impact on the wording of the items. The items responses reflect an amount of a SEL construct and Rosen, Glennie, Dalton, Lennon, and Bozick (2010) noted that if an item in the construct is modified (e.g., change of item wording or item is deleted/added from a previously published instrument) then the meaning of the construct may change. Often, items are administered in surveys

completed by participants where they endorse the item(s) (e.g. strongly agree the description of the item is like me). The items are ordinal in scaling (e.g., ordered categories), but as Agresti (2013) described, the distance between categories are unknown. Items with ordinal scaling are also known as polytomous indicators and different estimation methods are needed than when estimating indicators with continuous scaling (e.g., weight). Since there is a possibility that SEL measures can be useful in narrowing the achievement gap, there is an urgency to set a protocol in how we use psychometric models to measure SEL constructs.

Statement of the Problem

The original intention of this study was to search for a plausible psychometric model to measure change in SEL outcomes with the same items and different participants across time. However, after a deeper review of the literature there did not appear to be coherent methods in measuring a multidimensional SEL construct for one time point let alone multiple time points. There is one technical report (Rodriguez, 2017) that used psychometric models to construct multiple unidimensional SEL constructs. There were many reports that inaccurately suggested an SEL instrument was appropriate to use once there is a reported reliability and validity value (Cox, Foster, Bamat, 2019; Kendziora & Yoder, 2016) or suggest using the same SEL measures across grades 3 to 12 without tailoring the items to reflect the developmental stage of the student (Meyer, Wang, & Rice, 2018). In research studies that used psychometric models to measure a multidimensional SEL construct, there was a pattern of applying the most complex model over the less complex model while

disregarding the framework of the SEL construct that is needed to supply context and support score interpretation and use.

Significance of the Study

Instead of following the path of solely relying on statistics to use the most complex model/least restricted model (e.g., using a bifactor model versus a two factor model) as the internal structure of a multidimensional SEL construct (for short called SEL construct), in this study, I propose building a validity argument for SEL constructs and making the case that the plausible SEL internal structure must be consistent with the IUA and checked to have the measurement invariance (MI) property. Validity is the extent to which evidence and theory support the interpretation and use of scores (AERA, APA, & NCME, 2014). As Millsap (2011) points out, in order to have meaningful SEL score comparisons across groups (e.g., ethnic community groups) the SEL internal structure needs to have the MI property. The establishment of a strong IUA, including evidence regarding the MI property, in an SEL construct can lead to systematic and meaningful methods for measuring SEL measures. Other forms of validity evidence (associations with other variables and fairness) will be covered. The scope of the study will be cross-sectional and a longitudinal study will be left for a future study.

Organization of the Study

The organization of this study consists of three sequential components. The first component is to describe the extent to which theory supports a proposed SEL internal structure for score interpretation and use. This pertains to the measures proposed in the framework - the qualitative definition of the SEL construct.

The second component is to evaluate evidence that supports the dimensionality of the SEL measure through a plausible internal structure for score interpretation and use – the measurement model of the SEL construct. This involves presenting plausible competing hypotheses about the internal structure and choosing the best fitting model for the internal structure for the SEL measure. The third component of this study pertains to MI, requiring the best fitting model to have the MI property. The MI property is required if the SEL construct is used to compare scores across groups.

Need for study

As there is no consistent method of measuring multidimensional SEL constructs, this study can serve as an example of using theory and psychometric models to measure multidimensional SEL constructs and to verify that the SEL constructs are structured the same across groups (property of MI). In addition, I address ways to find a model that has the MI property that supports the use of the SEL construct to meaningfully compare groups. These proposed methods may also extend to longitudinal studies.

Research questions

1. Based on theory and evidence, what are the plausible competing models for the internal structure of selected SEL constructs?

   (a) Following the example from the technical report (Rodriguez, 2017), the evidence includes three pieces of information coming from confirmatory factor analysis (CFA) models:

      a. Model fit of the internal SEL structure
      b. Item loadings onto the SEL factors
      c. Correlations among the factors

   (b) Item Response Theory (IRT) models will be conducted to estimate the item and person location on a latent trait continuum (unidimensional) or multiple latent trait continuum (multidimensional).

2. Does the best fitting model hold the MI property needed to meaningfully compare scores across racial/ethnic community groups? Briefly see if the domain specific factors predict Grades when holding the general factor constant.
3. If the MI property is not attainable for the model in research question 2, is this due to:

   (a) the magnitude of the effect of individual idiosyncratic response styles?

   (b) incorrect or overly-complex model selection?

Chapter 2 contains a literature review of SEL theories and frameworks, connection between classical test theory (CTT) and IRT, and plausible psychometric models from CTT and IRT. The plausible models include: factor analysis, unidimensional IRT (UIRT), and multidimensional IRT (MIRT). Chapter 3 contains the methods section. Chapter 4 contains the results of the analysis. Finally, chapter 5 contains the implications of the results to the psychometric field and SEL community.

Chapter II: Literature Review

Introduction

Kyllonen and Zu (2019) noted there are over 270 SEL constructs pertaining to many frameworks, so the intention is not to cover all frameworks nor all the measures in the SEL construct, but focus on some. Although there are many frameworks to choose from, it is not common to use psychometric models for SEL measurement. The psychometric models presented are plausibly used as measurement models for SEL constructs.

Frameworks

CASEL

The CASEL (2015) framework introduced SEL as a process to develop social and emotional competence so children and youth can put into use the knowledge, attitudes, and skills necessary to regulate emotions, set and achieve goals, feel and show empathy for others, and make responsible decisions (p. 5). Furthermore, student learning takes place in multiple places in the child's and youth's environment: classrooms, schools, and family/community partnerships. All these three environments have the potential of supportive relationships that help learning.

21st Century Skills

There are several frameworks that describe 21st century skills (De Fruyt, Willie, & John, 2015). The Education Reform (2016, August 25) states 21st century skills include knowledge, skills (e.g., communication skills, problem solving skills),

work habits (e.g., adaptability, self-management, self-development, time-management), and character traits that are important for the person to be successful in a workplace. In the education setting, 21st century skills are skills that prepare students for the future (Care, Griffin, Wilson, 2018), and the Assessment and Teaching of 21st Century Skills (Griffin, McGaw, & Care, 2012) project focused on two skills: Collaborative problem-solving and Information and Communication Technologies (ICT) Literacy in digital networks. Griffin et al. (2012) defined collaborative problem-solving as: being able to see the perspective of other classmates in a group (e.g., perspective taking); positively engage as a member of a group (e.g., interpersonal skills such as sharing of knowledge, experience and expertise); collect and manage contributions (e.g., task regulation); take initiative and adopt a procedure to resolving a problem (e.g., social regulation); strive to build and develop knowledge and understanding in the group (e.g., knowledge building). The ICT Literacy in digital networks consists of the student's navigation of technology that prepares them for the future (Griffin, McGraw, & Care, 2012).

Big Five from Personality Psychology

Personality characteristics have the connotation of character or virtues that are related more with the person's temperament characteristics (Shiner & Caspi, 2003). The Big Five framework are personality characteristics consisting of conscientiousness, extraversion, agreeableness, emotional stability (also known as neuroticism), and openness to experience (Chernyshenko, Kankaraš, & Drasgow, 2018). Conscientiousness is about being self-disciplined, diligent, dependable, and

goal oriented. Extraversion refers to being socially connected, assertive, comfortable providing leadership, and passionate. Extraversion is more about the pursuit of interactions with others and agreeableness is about how well people get along with others. Agreeableness refers to being well-liked, respected and sensitive to the needs of others. The degree of being able to control emotional responses and mood is emotional stability. Openness to experience represents curiosity, excitement of creating something new, appreciation of beauty in the world, and self-reflection. Personality traits tend to be consistent characteristics of an individual in everyday situations but are prone to change in adulthood (Roberts & Mroczek, 2008).

Positive Youth Development

Positive Youth Development (PYD) is born from positive psychology (Lerner et al., 2005). Martin Seligman (2000), founder of positive psychology, changed the focus in the field of psychology from reducing mental illness (e.g., depression) to increasing positive emotion (e.g., happiness), building on strengths (e.g. productive positive traits) and having a life with meaning (e.g., using strengths for the greater good). This perspective can be applied to the general public since most people want more positive emotion, to build on strengths and to live a life with meaning. Seligman's focus on the well-being of an individual in positive psychology was applied to youth, as PYD focuses on the strengths of the adolescent and their development of assets (Lerner et al., 2005).

Character Strengths

The Character Strengths framework (Character Lab, 2017) pertains to strengths of heart, mind, and will. Strength of mind includes creativity (open to new ideas and possibilities), curiosity (wanting to know more), growth mindset, and intellectual humility (acknowledging limitations of one's knowledge). Strength of heart represents emotional intelligence (understanding and comfortable expressing feelings), honesty, gratitude, purpose, kindness, and social intelligence (ability to connect with people). Strength of will includes grit (perseverance), proactivity (taking initiative), and self-control

Measurement Models

From the paradigm for measuring SEL constructs (Figure 1.1), introduced in Chapter 1, depicts the framework (designed by researchers) as the qualitative structure of the construct and the construct is operationalized through the measures that are composed of item responses (measurement model of construct). According to Wilson (2005), the participant's position on the construct *causes* the responses to the items (the building blocks of a given measure); the observed responses are scored (e.g., 1= *strongly disagree*, 2 = *disagree*, 3 = *agree*, 4 = *strongly agree*) in an outcome space; plausible measurement models are applied to analyze the scored responses and to enable an inference to the underlying continuum of the construct. Notice the measurement model is applied at the measure-level and the results are inferred at the construct-level. The measurement models consist of psychometric models from two different approaches: CFA and IRT.

Before presenting the CFA approach, the differences between IRT and classical test theory (CTT) are presented. In general, the CTT approach focuses on a single score model using total scores and the IRT approach includes multiple score models using separate model equations for each item (Albano, 2020). An example of a CTT model is a raw-score model (e.g., the total sum of number of correct responses) where all items are equally weighted. The participant's expected total score in CTT is an estimate of the participant's true score that represents the person's ability (Hambleton, Swaminathan & Rogers, 1991). However, in CTT the person's ability depends on the selection of items in the assessment and the item properties depend on the group of participants who completed the assessment. In contrast, "an IRT model specifies how both trait level [person's ability] and item properties are related to the [participant's] item responses" (Embretson & Reise, 2000, p. 40).

The CFA and IRT models are equivalent when the latent trait is unidimensional and composed of dichotomous items. Furthermore, the latent variables in CFA are composed of items responses and each item is connected to the latent variable through a regression line.

Association Between Regression and Confirmatory Factor Analysis

Unidimensional CFA (e.g., a single latent variable) employs the correlations among the items with the purpose to explain the items' covariances through a pre-specified single factor structure (Brown, 2015). The CFA models for continuous items contain factor loadings, unique variances, and factor variance (and factor covariances for multidimensional models), which, for sample data, are all parameter estimates. In

certain CFA multidimensional models (e.g., correlated traits), each item loads on only one factor. For example, each loading for a continuous item on a factor is a simple linear regression line:

$$y_i = \beta_0 + \beta_1 X_i + \varepsilon_1, \quad (2.1)$$

where y_i (criterion variables) are the indicators (the observed response to the items in the survey), β_0 is the indicator intercepts (item conditional means) which can be set to zero in CFA (standardized model, mean at zero and variance at one), β_1 are the slopes which are the factor loadings of the indicators, X_i are the factor scores, and ε_1 are the residuals (the difference between the observed y and predicted y). The measurement equation with only one indicator (that is a continuous item) can be written similarly to equation 2.1 as:

$$y_1 = \nu_1 + \lambda_{11}\eta_1 + \varepsilon_1, \quad (2.2)$$

where y_1 is the measured variable (item 1), ν_1 is the measurement intercept for one indicator (fixed to zero), λ_{11} is the loading of the first item on factor one (fixed to 1.00 for the unit-loading identification), η_1 is the factor score, and ε_1 is the error for item 1. The CFA model parallels the CTT framework, where if the factor loadings in CFA are set to 1 then true-score variance is similar to the CTT framework.

In the CTT equation, the variance of the observed score is partitioned into the variance due to true score and variance due to measurement error (Novick, 1966):

$$\mathrm{Var}(X) = \mathrm{Var}(T) + \mathrm{Var}(E) \quad (2.3)$$

There are four assumptions in equation 2.3. First, the expected value of the observed score is the true score, so $E(X) = T$. Second, the errors are independent and true score are independent. Third, the errors in one assessment and the errors from a different assessment are independent. Lastly, the errors on one assessment and the true score on another assessment are independent. Similarly in CFA, the variance of a single indicator (observed response) is partitioned into the variance due to the factor variable and variance due to measurement error (Newsom & Ebrary, 2015) similar to equation 2.3:

$$\mathrm{Var}(y_1) = {\lambda_{11}}^2 * \mathrm{Var}(\eta_1) + \mathrm{Var}(\varepsilon_1). \quad (2.4)$$

Recall that the factor loading for item 1 in equation 2.2, λ_{11}, is fixed to 1.00. In a CFA model where the measurement errors are uncorrelated (Jöreskog, 1971), the proportion of the indicator's variance (y_1) that is explained by the factor (η_1) is the indicator's reliability, R^2 or the reliability estimate (ρ):

$$R^2 = \frac{\mathrm{Var}(T)}{\mathrm{Var}(X)} = \rho = \frac{{\lambda_{1i}}^2 * \mathrm{Var}(\eta_1)}{\mathrm{Var}(y_1)}, \quad (2.5)$$

where ρ is the proportion of variance (where ${\lambda_{1i}}^2$ does not have to equal 1) in the indicator that is explained by the factor variance, $\mathrm{Var}(\eta_1)$. In CFA, there is usually more than one indicator and the factor loadings can be any positive or negative value. The observed response of the indicators in a CFA form a population variance-covariance matrix.

The structural equation measurement model for a CFA model is (Newsom, 2015):

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}_{\boldsymbol{\varepsilon}}, \quad (2.6)$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is a model-implied variance-covariance matrix, $\boldsymbol{\theta}$ are the free parameters in a hypothesized model, $\boldsymbol{\Lambda}$ is the factor loading matrix, $\boldsymbol{\Psi}$ is the factor variance matrix, $\boldsymbol{\Lambda}'$ the transpose of the factor loading matrix, and $\boldsymbol{\Theta}_{\boldsymbol{\varepsilon}}$ is the measurement error matrix. In CFA, we test the hypothesis that the population variance-covariance matrix ($\boldsymbol{\Sigma}$) of observed variables is equal to the variance-covariance matrix implied by a hypothesized model $\boldsymbol{\Sigma}(\boldsymbol{\theta})$:

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}), \quad (2.7)$$

where $\boldsymbol{\Sigma}$ (population variance-covariance matrix of observed variables) is a function of the free parameters (θ) in a hypothesized model. Similar to regression models, results of the CFA models are presented at the model-level (e.g., coefficient of determination/reliability, equation 2.5) and coefficient level (e.g., slopes of the items, λ presented in equation 2.4).

Model Estimation. The purpose of the CFA model is to find a set of parameters θ to produce $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ so that the difference between $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ can be minimized (Wang & Wang, 2012). This is somewhat similar to minimizing the discrepancies between fitted and observed values of the outcome variable in a regression model $\Sigma(Y - \widehat{Y})$. In CFA, the estimation procedure minimizes the residuals which are the discrepancies between $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$**.** Both the population variance-covariance matrix ($\boldsymbol{\Sigma}$) and model-implied variance-covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, are unknown so $\left[\mathbf{S} - \boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right)\right]$ or $(S - \widehat{\boldsymbol{\Sigma}})$ is used to minimize the discrepancies between the

model-estimated/implied variance-covariance matrix $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ and the sample variance-covariance matrix (**S**). Note the model parameter estimates are noted by $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ can also be noted as $\widehat{\boldsymbol{\Sigma}}$**.**

The estimation function commonly used for continuous indicators is maximum likelihood (ML; Bollen, 1989) and for ordered categorical indicators is the weighted least squares (WLS) based robust estimators, specifically the mean and variance-adjusted WLS (WLSMV). In *Mplus 8.3* (Muthén & Muthén, 2019) the WLSMV estimators use the probit link.

Fit Indices. The CFA models are evaluated on the overall goodness-of-fit and local fit indices. The overall goodness-of-fit indices is the chi-square statistic (χ^2). To assess how close the model-implied variance-covariance matrix ($\boldsymbol{\Sigma}(\boldsymbol{\theta})$) is to the population covariance-variance matrix (**Σ**) the following indices are reported**:** standardized root mean square residual (SRMR), root mean square of approximation (RMSEA), comparative fit index (CFI) and Tucker-Lewis index (TLI). These overall goodness-of-fit indices provide a global fit of the model (Brown, 2015), .

The chi-square statistic (χ^2) is, an absolute fit measure, known as the chi-square test that assesses the null hypothesis (H_0). The null hypothesis states the population variance-covariance matrix (**Σ**) is equal to the model-implied variance-covariance matrix ($\boldsymbol{\Sigma}(\boldsymbol{\theta})$), presented in equation 2.7. In other words, the null hypothesis states that our specified model (represented by the model-implied variance-covariance matrix [$\boldsymbol{\Sigma}(\boldsymbol{\theta})$]) fits the data just as well as the fully saturated model (population variance-covariance matrix (**Σ**)).

The desired outcome is for the chi-square statistic to have a *p*-value *larger* than .05 (the data are consistent with the null hypothesis, therefore, the null hypothesis is not rejected); this suggests the specified CFA model supports the plausibility of the postulated associations among the items (Wang & Wang, 2012). The null hypothesis can also imply the residual matrix is zero (H_0: residual matrix $= 0$), and if the null is not rejected then this indicates there is no difference between the model-estimated/implied variance-covariance matrix $\widehat{\boldsymbol{\Sigma}}$ and the sample variance-covariance matrix ($\mathbf{S}$). Conversely, if the *p*-value is less than .05, then the fully-saturated model fits better than our specified model (rejecting the null hypothesis to favor the alternative hypothesis). A *p*-value smaller than .05 indicates the data are *not* consistent with the null hypothesis which infers that the residual matrix is *not* zero ($\mathbf{S} - \widehat{\boldsymbol{\Sigma}} \neq \mathbf{0}$).

The chi-square test should not be the sole reason to evaluate (e.g., reject) a model, as there are other model indices needed to evaluate the model. The SRMR is the residual-based absolute fit index, where values below .08 are considered a good fit (Hu and Bentler, 1999) and less than .10 are acceptable (Kline, 2016). The error of approximation indicates the lack of fit between the specified model to the population measured in RMSEA where values less than .06 are interpreted as a good fit (Hu and Bentler, 1999). Comparative (incremental) fit indices include CFI and TLI where good fit values are closer to 1.00. TLI has a penalty for model complexity (more free parameters in the specified model) and a value less than .90 represents the need to respecify the model (Wang & Wang, 2012).

Nested models need to be estimated from the same dataset, usually the likelihood ratio (LR) test is used to compare models. The chi-square difference test is used for discriminant validity where the null hypothesis states there is no difference in model fit between the more complex (less restricted) model and the less complex (more restricted) model. However, if the *p*-value is less than .05 then the alternative hypothesis is applied that says the more complex model fits better than the less complex model. Usually the complex model is preferred than the simple model, so the desired outcome is for the *p*-value to be smaller than .05. Since the WLSMV is used for estimation of parameters with polytomous items, LR test cannot be used since the nested models do not follow a χ^2 distribution.

Other criterion statistics include Akaike's information criterion (AIC), Bayesian information criterion (BIC) and -2loglikelihood (-2LL) where the lower values indicate a better fitting model. These information criterion indices are used for model comparisons since the AIC, BIC and -2LL are relative fit statistics (Wang & Wang, 2012). An additional measure of overall fit is Goodness of Fit Index (GFI) proposed by Jöreskog and Sörbom in 1986 (according to Bollen, 1989), which uses ML to measure the fit between the model-implied variance-covariance matrix ($\boldsymbol{\Sigma}(\boldsymbol{\theta})$), and population variance-covariance matrix (**Σ).**

Although not part of the model indices, the standardized residuals are a way to assess the how close the model-estimated/implied variance-covariance matrix $\widehat{\boldsymbol{\Sigma}}$ to the sample variance-covariance matrix (**S**). Small standardized residuals are desired. Large standardized values, larger than 2.58 in magnitude according to Jöreskog and

Sörbom (1989), “indicate a large discrepancy in a specific variance or covariance between **S** and $\widehat{\boldsymbol{\Sigma}}$” (Wang & Wang, 2012, p. 23).

As stated earlier the in a standard CFA model, after the covariances of the indicator variables is accounted for by an underlying latent variable then these indicator variables should not be correlated with each other (Wang & Wang, 2020). This is also known as the local independence assumption. In CFA, the item residuals are set to zero but sometimes the modification indices given at the end of the software program output, specify the item residuals to be correlated which states the covariance in the item indicators are explained by the underlying factor and item residual as well as something else outside the CFA model. There needs to be a reason for correlating item residuals (e.g., measurement errors correlated in a longitudinal study) other than improving the model fit.

Model Identification. The CFA model needs to be over-identified (there are fewer parameters estimated in the model than number of variances and covariances of the indicators) to have a unique solution for the free parameters in the hypothesized model (Wang & Wang, 2012). The equation to find the number of variances and covariances (elements or pieces of observed information) of the input is:

$$\text{elements} = a = \frac{p(p+1)}{2},$$

where p is the total number of observed variables (number of items).

To find the degrees of freedom (*df*):

$$df = a - \text{free parameters},$$

where a is the elements and the free parameters are the number of freely estimated parameters. These free parameters may include (first-order and second-order) estimation of factor loadings, indicator variances, and factor variances. When the df is positive (there are more elements than freely estimated parameters), the model is over-identified. When *df* is negative (more freely estimated parameters than elements), the model is under-identified, and when it is zero (same amount of elements as freely estimated parameters), the model is just-identified. The counting rule states that there needs to be more elements than freely estimated parameters for the model to be identified, this means the model has a set of unique estimates (Kline, 2016). The under-identified model has an infinite set of solutions for the estimates and the just-identified model can perfectly reproduce the sample covariance matrix but may not be consistent with a certain hypothesis.

Confirmatory Factor Analysis Models

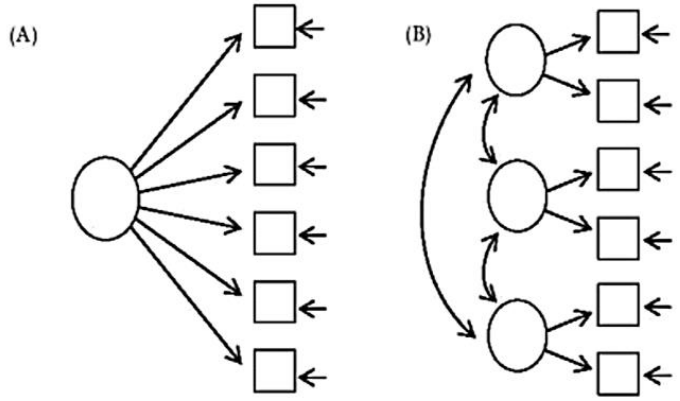
In a CFA model the factors are theoretically defined. The framework-level qualitative definitions are used to base the dimensionality of SEL constructs. The CFA models evaluate the extent to which a set of items (indicator variables) in an assessment measures the latent variables they are theorized to measure. In other words, "the latent variables are hypothesized to explain the covariances in the observed indicator variables" (Wang & Wang, 2012, p. 35). The simplest CFA model has one factor (latent variable) and all the items load onto one factor where the item residuals are not correlated with each other (see Figure 2.1A).

Correlated Traits. The correlated traits model has more than one factor (latent variable/trait) and there are associations between the factors (see Figure 2.1B). Similar to the one factor model, the items load onto only one factor and the item residuals are uncorrelated with each other. The CFA model can be expressed as (Wang & Wang, 2012):

$$X = \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}, \tag{2.8}$$

Where X contains the observed scores of the indicator variables, matrix $\boldsymbol{\Lambda}_x$ contains the factor loadings; $\boldsymbol{\xi}$ contains the factor scores; and $\boldsymbol{\delta}$ contains the residuals. Figure 2.1B has three correlated latent variables with two items loading onto each latent variable (and item residuals are uncorrelated with each other).

Figure 2.1
(A) Unidimensional Model and (B) Correlated Traits Model



Note. Adapted from Reise and Revicki (2015)

Second Order. A second-order model has a hierarchical structure where each item loads on one first-order factor and the first-order factors load on the higher-order factor also known as the common factor (Brown, 2015). Using the second-order model is advantageous when the lower-order factors are significantly intercorrelated since the

hypothesized common factor accounts for the correlations of the lower-order factors (Chen, Sousa & West, 2005).

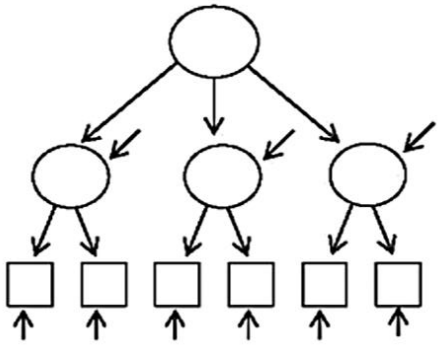
The second-order equations, equation 2.8, can be expressed as (Chen, West & Sousa, 2006):

$$\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \quad \text{eq. 1}$$

$$\boldsymbol{Y} = \boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{\eta} + \boldsymbol{\epsilon} \quad \text{eq. 2} \tag{2.8}$$

Equation 1 is for the common factor structure and equation 2 is for the measurement model that contain the lower-order factors. In equation 2, *Y* contains the indicator variables; matrix Λ_y contains the item loadings onto the lower-order factors; $\boldsymbol{\eta}$ contains the lower-level factors; and $\boldsymbol{\epsilon}$ contains the measurement errors of *y*. For equation 1, vector $\boldsymbol{\Gamma}$ contains the lower-order factors loadings onto the common factor; vector $\boldsymbol{\xi}$ contains the common factor; vector $\boldsymbol{\zeta}$ contains the disturbances of the lower-order factors (these are unique variances that are not accounted for in the common factor). Figure 2.2 is an example of a second-order model with six indicator variables, each loading onto one lower-order factor, and the three lower-order factors loading onto one common factor.

Figure 2.2
Second-Order model with One Common Factor and three Lower-Level Factors



Note. Adapted from Reise and Revicki (2015)

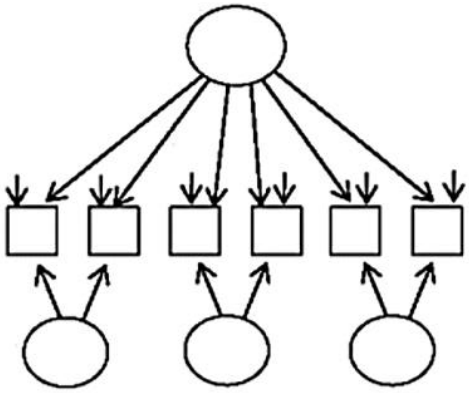
Bifactor. In a bifactor model, each item loads first on a general factor and then onto one domain-specific factor where the direct effects between specific domain factors and items can be evaluated (Schmid & Leiman, 1957). The general and domain specific factors are orthogonal (Brown, 2015). The bifactor model was first introduced by Holzinger and Swineford (1937) and is used in psychological measures (Brouwer, Meijer, Weekers, & Baneke 2008; Lahey et al., 2012; Olatunji et al., 2019; Osman et al., 2010; Patrick, Hicks, Nichol, & Krueger, 2007), and testlets in assessments (DeMars, 2006), but not applied in education for SEL measures. The bifactor equation can be expressed as (Chen, West & Sousa, 2006):

$$Y = \boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{2.9}$$

where vector *Y* contains the indicator variables; matrix $\boldsymbol{\Lambda}_{\boldsymbol{y}}$ contains the factor loadings of the general and domain specific factors; vector $\boldsymbol{\eta}$ contains the general and domain specific factors; and $\boldsymbol{\epsilon}$ contains the residuals of *y*. A bifactor model has the advantage over a second-order model in being able to separate the general and specific factors to predict an external criterion variable. Figure 2.3 is an example of six indicator

variables loading onto the general factor and then to one of the domain-specific factors where the general factor and the domain-specific factors are orthogonal to each other. Notice that if the loadings in the general factor are fixed to zero and the specific domain factors are correlated, then the model becomes a correlated traits model (Figure 2.1B). Also note, setting the specific domain factor loadings to zero in the bifactor model leaves the items to load on the general factor which would be similar to a one-factor model (Figure 2.1A).

Figure 2.3
Bifactor Model with One General Factor and Three Domain Specific Factors



Note. Adapted from Reise and Revicki (2015)

Two-Tier. The bifactor factor and two-tier models both have domain specific factors. Unlike the bifactor factor that has one general factor, the two-tier has two primary (general) factors that are allowed to correlate with each other (but not with the domain specific factors). As the name implies, the two-tier factor has the primary factors in the first-tier and the domain specific factors in the second-tier (Reise & Revicki, 2015).

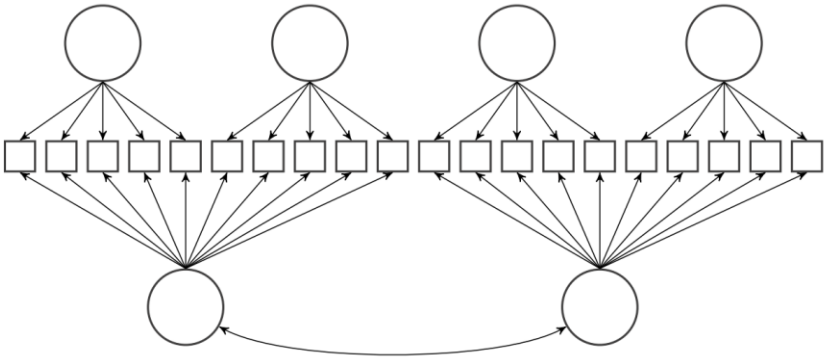
In the first tier, the items are free to load on the two primary factors – this makes the two-tier more flexible than the bifactor model. However, the item cross-loadings onto the domain-specific factors are not allowed in the second-tier. The domain specific factors are mutually orthogonal to each other and also uncorrelated with the primary factors (Cai, 2010). The standard two-tier factor equation can be expressed (Reise & Revicki, 2015)

$$Y = \boldsymbol{\Lambda_y \eta} + \boldsymbol{\epsilon}, \tag{2.10}$$

where vector *Y* contains the indicator variables; matrix $\boldsymbol{\Lambda_y}$ contains the factor loadings of the primary factors and domain specific factors; vector $\boldsymbol{\eta}$ contains the primary factors ($\boldsymbol{\eta}$) in the first-tier and domain specific factors ($\boldsymbol{\xi}$) in the second-tier; and $\boldsymbol{\epsilon}$ contains the residuals. Although the standard two-tier measurement model (equation 2.10) is identical to the bifactor measurement model (equation 2.9), the two-tier factor model includes multiple primary factors in tier-one (that are correlated with each other) and specific domain factors in tier-two (that are orthogonal to each other and the primary factors). For example, the two-tier model in Figure 2.4 has two correlated primary factors and four domain specific factors that are orthogonal to each other as well as to the primary factors. Note that the items load onto one of the primary factors as well as on one of the four domain specific factors.

Figure 2.4

Two-Tier model with Two Primary and Four Domain Specific Factors



Note. Adapted from Cai (2010).

The two-tier model can also be used to examine individual idiosyncrasies in response style by adding a random intercept factor which is an extension of Maydeu-Olivares and Coffman (2006) item factor analysis model. An example of this two-tier model would include a general factor, domain-specific factors, and random intercept factor.

Association Between CFA and IRT

The CFA with binary outcomes (e.g., *yes* or *no* response options) is equivalent to a two-parameter (2PL) normal ogive IRT model (Glockner-Rist & Hoijtink, 2003). In other words, the CFA approach is equivalent to the IRT approach when the items are dichotomous. Recall that from the CFA approach the equation for a one-factor continuous model with one item can be written:

$$y_1 = \mathrm{v}_1 + \lambda_{11}\xi_1 + \varepsilon_1, \tag{2.11}$$

In CFA, when the factor model is not continuous but binary, the one-factor model for the latent response variable (y_i^*) can be written as (Kamata & Bauer, 2008):

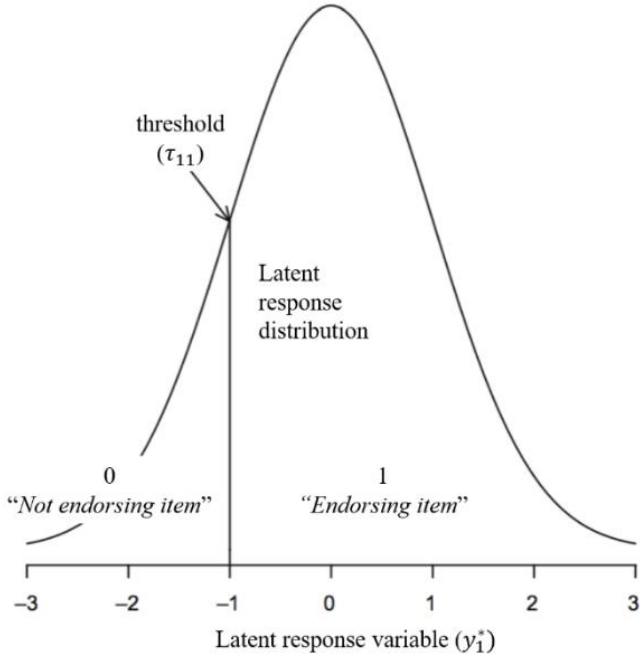
$$y_i^* = v_i + \lambda_i \xi + \varepsilon_i \,, \tag{2.12}$$

where v_i is the intercept, λ_i are the factor loadings, ξ is the latent variable (continuous factor score), and ε_i is the residual for item *i*. The observed response (y_i) is not continuous but dichotomous (binary), so the response to an item has two possible outcomes (e.g., 1 = *yes* and 0 = *no*) and can be expressed as:

$$y_i = \begin{cases} 1, \text{if } y_i^* \geq \tau_i \\ 0, \text{if } y_i^* < \tau_i \end{cases} \tag{2.13}$$

where τ_i is the threshold for item *i* and for a dichotomous item there is only one threshold, $y_i^*$is the latent response variable (LRV). The distribution of the LRV (y_i^*) can be visually seen in Figure 2.5 for one item, where the observed response (y_i) of 1 is *endorsing the item* and 0 is *not endorsing the item*. Notice the LRV is continuous (from negative 3 to positive 3, in Figure 2.5) as the latent response distribution is normally distributed and the threshold (τ_{11}) separates the responses from not endorsing the item ($y_i = 0$) and endorsing the item ($y_i = 1$).

Figure 2.5
Latent Response Distribution for a Single Dichotomous Item 1

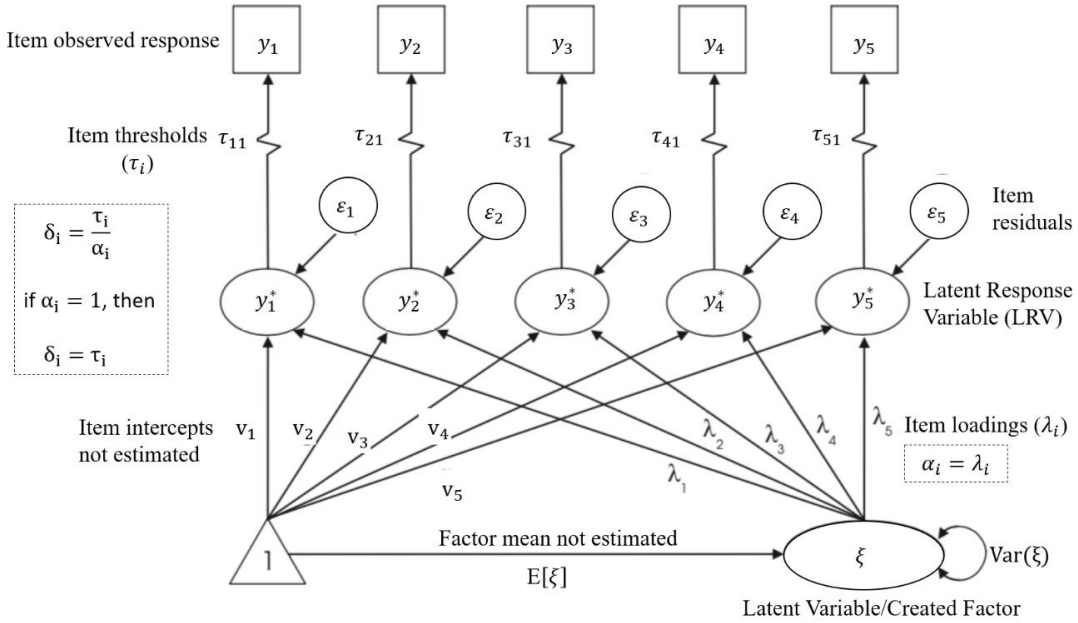


Note. The observed indicator (y_i) has two outcomes: 0 or 1. Adapted from Kline (2016).

To be able to associate a CFA model with an IRT model, the item intercepts and factor mean in the CFA model will not be estimated but instead assumed to be zero. Displayed in Figure 2.6, the LRV has a non-linear association with the indicators (items) but a linear association with the factor (Kline, 2016).

Figure 2.6

Single-factor CFA with Dichotomous Items/Indicators



Note. Factor is standardized and theta scaling used for LRV. Adopted from Kline (2016).

The latent variable (created factor) and LRV in Figure 2.6 need to be scaled. Kamata and Bauer (2008) explained the equivalence of the CFA model to an IRT model can be seen once the scaling of the factor (latent variable) and LRV were decided. The factor is commonly scaled as having the first loading be the reference indicator (fixing the first item loading to 1 and threshold to zero) or standardizing the factor (the factor mean and variance constrained where $\mathrm{E}[\xi] = 0$ and $\mathrm{Var}(\xi) = 1$).

The scaling of the LRV is known as Delta scaling (marginal parameterization of the LRV where the $\mathrm{Var}(y^*) = 1$) and Theta scaling (conditional parametrization of the LRV where the $\mathrm{Var}(\varepsilon_i) = 1$). If the factor is scaled to fix the reference indicator then the factor mean and variance are freely estimated. However, if the factor is standardized then the factor loadings and thresholds are freely estimated. The Delta

scaling for the LRV fixes the variance to 1 and estimates the error variance for the LRV. For Theta scaling, the error variance is fixed to 1 and the LRV variance is estimated. Table 2.1 shows the four options of scaling the factor and LRV: Option 1 is marginal-unit-loading identification parameterization, option 2 is marginal-standardized parameterization, option 3 is conditional-unit-loading identification parameterization, and option 4 is conditional-standardized parameterization.

Table 2.1

Scaling the Latent Response Variable and Factor

	Reference Indicator (factor) (Fix/Estimate)	Standardized Factor (Fix/Estimate)
	Option 1 Fix $\lambda_1 = 1, \tau_1 = 0$ Estimate $\mathrm{E}(\xi_1), \mathrm{Var}(\xi_1)$	*Option 2* Fix $\mathrm{E}(\xi_1) = 0, \mathrm{Var}(\xi_1) = 1$ Estimate λ_1, τ_1
Delta Scaling	$\mathrm{Var}(y^*) = 1$ $Var(\varepsilon_i) = 1 - \lambda_1^2 \mathrm{Var}(\xi_1)$ $Var(\varepsilon_i) = 1 - \mathrm{Var}(\xi_1)$	$\mathrm{Var}(y^*) = 1$ $\mathrm{Var}(\varepsilon_i) = 1 - \lambda_1^2 \mathrm{Var}(\xi_1)$ $\mathrm{Var}(\varepsilon_i) = 1 - \lambda_1^2$
	Option 3 Fix $\lambda_1 = 1, \tau_1 = 0$ Estimate $\mathrm{E}(\xi_1), \mathrm{Var}(\xi_1)$	*Option 4* Fix $\mathrm{E}(\xi_1) = 0, \mathrm{Var}(\xi_1) = 1$ Estimate λ_1, τ_1
Theta Scaling	$\mathrm{Var}(\varepsilon_i) = 1$ $\mathrm{Var}(y^*) = \lambda_1^2 \mathrm{Var}(\xi_1) + 1$ $\mathrm{Var}(y^*) = \mathrm{Var}(\xi_1) + 1$	$\mathrm{Var}(\varepsilon_i) = 1$ $\mathrm{Var}(y^*) = \lambda_1^2 \mathrm{Var}(\xi_1) + 1$ $\mathrm{Var}(y^*) = \lambda_1^2 + 1$

Note. Delta scaling uses the marginal distribution and theta scaling uses the conditional distribution.

In Kamata and Bauer (2008) the 2PL IRT model is expressed as:

$$p_i(y_i = 1|\xi) = f(\alpha_i\xi + \beta_i), \tag{2.14}$$

where the $p_i(y_i = 1|\xi)$ is the probability of an observed response of 1 given the person's factor score, α_i is the slope parameter, β_i is the intercept parameter (previously referred as v_i) for item *i*, ξ is the latent score for a certain person, and *f* is a cumulative distribution function (CDF). According to Takane and Leeuw (1987), when the factor is standardized and the LRV uses the conditional distribution (Option 4 in Table 2.1), then the parameters for a 2PL IRT unidimensional model can be expressed as:

$$\alpha_i = \frac{\lambda_i}{\mathrm{Var}(\varepsilon_i)^{\frac{1}{2}}}, \tag{2.15}$$

$$\beta_i = \frac{-\tau_i}{\mathrm{Var}(\varepsilon_i)^{\frac{1}{2}}}, \tag{2.16}$$

where α_i is the item discrimination parameter, λ_i is the loading for item *i*, and $\mathrm{Var}(\varepsilon_i)^{\frac{1}{2}}$ is the square root of the error variance for the LRV, β_i the intercept parameter for item *i*, and τ_i is the threshold for item *i*. Recall that in option 4 from Table 2.1, the error variance of the LRV is fixed to 1, so substituting $\mathrm{Var}(\varepsilon_i) = 1$ into equation 2.15 results in $\alpha_i = \lambda_i$. The IRT item discrimination parameter (α_i) is equivalent to the factor analysis (FA) item loading parameter (λ_i). Also, substituting $\mathrm{Var}(\varepsilon_i) = 1$ into equation 2.16 leads to $\beta_i = -\tau_i$ where the FA item intercept (β_i) is equal to the negative FA threshold ($-\tau_i$).

The equivalence between the FA threshold and IRT item difficulty is shown when the 2PL IRT model is expressed as (Kamata & Bauer, 2008):

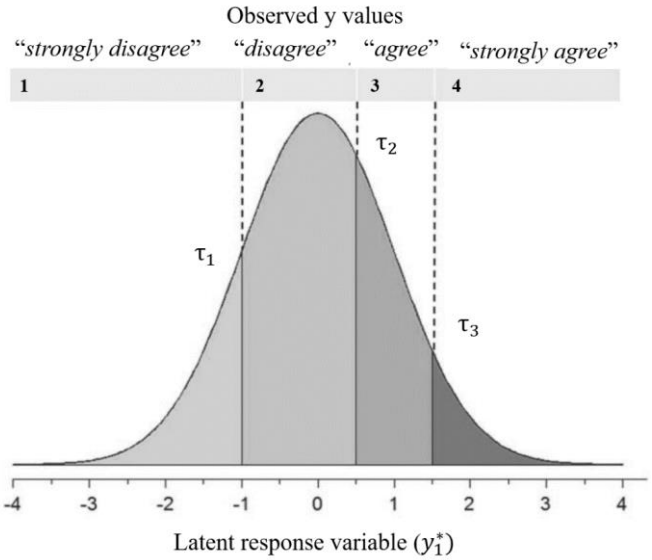
$$f[\alpha_i(\xi - \delta_i)], \tag{2.17}$$

where α_i is the item discrimination parameter, ξ is the person's factor score (person's ability), δ_i is the item difficulty (location) and *f* is a CDF. Using equation 2.17 from the IRT approach, the $\delta_i = -\beta_i/\alpha_i$ and the FA factor intercept $\beta_i = -\tau_i$ is substituted into $\delta_i = -\beta_i/\alpha_i$ then $\delta_i = \tau_i/\alpha_i$. In the Rasch model the item discrimination is constrained to 1 (α_i=1), so the item difficulty is equivalent to the item threshold ($\delta_i = \tau_i$). When the factor is standardized and the LRV uses the conditional distribution (theta scaling) then the item discrimination is equivalent to the item loadings ($\alpha_i = \lambda_i$). Furthermore, if the item discrimination is fixed to 1 then the item difficulty (location) in an 2PL IRT model is the same as the item threshold in a dichotomous FA model ($\delta_i = \tau_i$). Both $\alpha_i = \lambda_i$ and $\delta_i = \tau_i$) are showed in Figure 2.6. Note that de Ayala (2009) expressed the IRT item intercept as $\gamma_i = -\delta_i\alpha_i$, so $\delta_i = -\gamma_i/\alpha_i$, equivalent to $\delta_i = -\beta_i/\alpha_i$.

For a polytomous item there is more than one threshold. Figure 2.7 visually displays a latent response distribution for a single ordinal-categorical item showing the observed response option distribution to a LRV. The item has four categorical levels (1 = *strongly disagree*, 2 = *disagree*, 3 = *agree*, 4 = *strongly agree*), three thresholds (τ_1, τ_2, τ_3) and the x-axis is the latent response variable for item 1.

Figure 2.7

Latent Continuous Responses and Observed Ordinal Responses. There are Four Response Options (1, 2, 3, 4) and Three Thresholds (τ_1, τ_2, τ_3).



Note. Adopted from Liu et al. (2017).

The association between a latent response distribution (y^*), and observed responses (y) can be written as (Flora & Curran, 2004):

$$y = c, \ \tau_c < y^* < \tau_{c+1},$$

$$c = 1, 2, \ldots, C - 1 \text{ and } \tau_0 = -\infty, \ \tau_C = \infty,$$

where τ are the thresholds , and c are the categories of the rating scale item (e.g., *strongly disagree* = 1 , *disagree* = 2, *agree* = 3, *strongly agree* = 4).

Item Response Theory Models

The purpose of a CFA model is to explain the covariances among the indicators and the latent response variable (y^*); however, with the limitation that it does not describe the probability of an individual endorsing an item (Bock, 1997). An IRT model accounts for participant's item responses since the IRT model connects the

characteristics of items (item parameters) and characteristics of persons (latent traits) to the probability of endorsing a particular response category (Bock, 1997). Item thresholds in the FA model translate to the item difficulty in the IRT model and the factor loadings in the FA are equivalent to the item discriminations in the IRT 2PL model. The item response curve for a 2PL model for dichotomous item responses is (Reise & Revicki, 2015):

$$\mathrm{p}\left(x_j = 1 \middle| \theta, \alpha_\mathrm{j}, \delta_\mathrm{j}\right) = \frac{\mathrm{e}^{-1.7\alpha_\mathrm{j}(\theta-\delta_\mathrm{j})}}{1+\mathrm{e}^{\alpha_\mathrm{j}(\theta-\delta_\mathrm{j})}}, \tag{2.18}$$

where α_j is the item discrimination, $\mathrm{p}\left(x_j = 1 \middle| \theta, \alpha_\mathrm{j}, \delta_\mathrm{j}\right)$ is the probability of the response of 1 (e.g., endorsing the item), α_i is an item discrimination and a function of the slope (factor loadings in CFA), δ_j is the item location (item thresholds in FA), θ is a continuous latent variable, and 1.7 is the scaling factor for the item slope parameter in logistic models to be equivalent to a normal-ogive model.

In IRT, θ is commonly known as the person parameter or ability; however, θ can also represent person location on a trait (the characteristic being measured). Person ability (θ) and item parameters (δ_j) are located on the same continuum. IRT models have the advantage of estimating person parameters (person's ability) and item parameters. There are a family of IRT models; however, only a few will be introduced in the order of least complex to most complex. The item information function provides the precision of estimates conditioned on θ, based on a given scoring formula.

Unidimensional models. The least complex IRT models are those with dichotomous items (e.g., *yes* or *no* item responses) and a single continuous latent variable (θ). The unidimensional model has a single ability θ (one underlying dimension ~ one factor in FA). Also, the response of the person is conditioned on the person's ability and not dependent on how the examinee responds to any other items (conditional local independence) and each IRT model expresses the interaction between item difficulty and person ability differently.

Rasch. The Rasch model, where the data fit the model, can be written as (de Ayala, 2009):

$$\mathrm{p}\left(x_j = 1 \middle| \theta, \delta_\mathrm{j}\right) = \frac{\mathrm{e}^{(\theta - \delta_j)}}{1 + \mathrm{e}^{(\theta - \delta_j)}}, \tag{2.19}$$

where $\alpha = 1$ is the item discrimination fixed at 1.00 for all items (factor loadings in the FA model fixed at 1.0), $\mathrm{p}\left(x_j = 1 \middle| \theta, \delta_j\right)$ is the probability of responding 1 to the item, θ is the person location parameter (continuous latent variable in FA), and δ_j is item *j*'s location (item thresholds in the FA model).

1PL. The next model is the one-parameter logistic (1PL) model (where the model fits the data) and the item discrimination is estimated to be the same for all the items (de Ayala, 2009):

$$\mathrm{p}\left(x_j = 1 \middle| \theta, \alpha, \delta_j\right) = \frac{\mathrm{e}^{\alpha(\theta - \delta_j)}}{1 + \mathrm{e}^{\alpha(\theta - \delta_j)}}, \tag{2.20}$$

where α is the item discrimination constrained to be the same for all items, $\mathrm{p}(x_j = 1|\theta, \alpha, \delta_j)$ is the probability of endorsing the item, θ is the person location parameter, and δ_j is item *j*'s location.

2PL. Next in complexity is the two-parameter logistic (2PL) where the item discrimination can vary in each item indicated by the subscript *j* (de Ayala, 2009):

$$\mathrm{p}(x_j = 1|\theta, \alpha_j, \delta_j) = \frac{\mathrm{e}^{\alpha_j(\theta-\delta_j)}}{1 + \mathrm{e}^{\alpha_j(\theta-\delta_j)}}, \tag{2.21}$$

where α_j is the item discrimination freely estimated for each item, $\mathrm{p}(x_j = 1|\theta, \alpha_j, \delta_j)$ is the probability of endorsing the item, θ is the person location parameter, and δ_j is the location of item j.

3PL. The three-parameter logistic (3PL) model, introduced by Birnbaum (1968), has a lower-asymptote parameter which may not apply to rating-scale items, but the equation is (de Ayala, 2009):

$$\mathrm{p}(x_j = 1|\theta, \alpha_j, \delta_j, \chi_j) = \chi_j + (1 - \chi_j)\frac{\mathrm{e}^{\alpha_j(\theta-\delta_j)}}{1+\mathrm{e}^{\alpha_j(\theta-\delta_j)}}, \tag{2.22}$$

where α_j is the item discrimination freely estimated for each item, $\mathrm{p}(x_j = 1|\theta, \alpha_j, \delta_j, \chi_j)$ is the probability of endorsing the item, θ is the person location parameter, and δ_j is item j's location, and χ_j is the lower-asymptote parameter. The 3PL model is technically no longer a logistic model as the lower asymptote not being 0 then the probability of a response of 1 at δ_j is no longer .50 (e.g., $\chi_\mathrm{j} = 0$, then $\frac{1+\chi_\mathrm{j}}{2} = .50$).

Partial Credit Model. The Rasch partial credit model gives each item an ordered response structure. The partial credit model takes into consideration the probability of polytomous items on a survey being endorsed. A Rasch partial credit model can be written as (Masters, 1982):

$$p(x_j|\theta,\delta_{jh}) = \frac{e^{\sum_{h=0}^{x_j}(\theta-\delta_{jh})}}{\sum_{k=0}^{m_j} e^{\sum_{h=0}^{k_j}(\theta-\delta_{jh})}}, \quad (2.23)$$

where $\sum_{k=0}^{0}(\theta-\delta_{jh}) \equiv 0$ (this does not indicate there is a zero threshold but set to zero to simplify the model), x_j is the score on item *j*, m_j is the maximum score on item *j*, θ is the person ability (latent trait), *k* is the response categories (e.g., *k* = 1, 2, …, , m_j), δ_{jh} is the transition location parameter for the *h*th score category for item *j*.

The transition parameter in a rating scale model was introduced by Andrich (1978), $\delta_{jh} = \delta_j + \tau_h$ that is the item location (δ_j) and item threshold (τ_h) for the rating scale items where the number of categories need to be the same for all items (de Ayala, 2009). There is one more response category than item thresholds and transition location parameters, for example, an item with four response categories (where *k* is 1 = *strongly*, 2 = *disagree,* 3 = *disagree, agree,* and 4 = *strongly disagree*) has three item thresholds (τ_1 is between *strongly disagree* and *disagree* categories, τ_2, is between *disagree* and *agree* categories and τ_3 is between *agree* and *strongly agree* categories) and three location transition parameters (e.g., for item 1, $\delta_{11} = \delta_1 + \tau_1$, $\delta_{12} = \delta_1 + \tau_2$, and $\delta_{13} = \delta_1 + \tau_3$). Andrich's rating scale model (1978) and Semejima's (1969) graded response model work well for rating items found in surveys

(Brennan, 2006), where the rating-scale structure (distance between thresholds) is fixed across items.

In partial credit items (e.g., a partial credit math problem) the transition location parameter is the relative difficulty of reaching level *h* in item *j*, so the transition location parameter specifies the difficulty of step from category *h*-1 ($x_j = 1$) to category h ($x_j = 2$). The number of categories, m_j, can vary by item *j*. As part of the Rasch model, the item discrimination is fixed to 1.00 for all items ($\alpha = 1$). Next is a non-Rasch model for ordered polytomous items.

Generalized Partial Credit. One non-Rasch model is the generalized partial credit (GPC) model that allows the item discriminations (α_j) to vary across items. A GPC model can be written (Muraki, 1992):

$$\mathrm{p}\left(x_{jk}\middle|\theta,\alpha_j,\delta_{jh}\right)=\frac{\mathrm{e}^{\sum_{\mathrm{h}=1}^{k_j}\alpha_{\mathrm{j}}(\theta-\delta_{jh})}}{\sum_{\mathrm{c}=1}^{m_j}\mathrm{e}^{\sum_{\mathrm{h}=1}^{\mathrm{c}}\alpha_{\mathrm{j}}(\theta-\delta_{jh})}}, \qquad 2.24$$

where $\mathrm{p}\left(x_{jk}\middle|\theta,\alpha_j,\delta_{jh}\right)$ is the probability of providing a response in item *j* with category k*th*, θ is the person ability, α_j is the item discrimination, δ_{jh}(where $\delta_{jh} = \delta_j + \tau_h$) is the transitional location parameter between the *h*th category and the *h*-1 category (for example, between category h ($x_j = 2$) and category *h*-1 ($x_j = 1$)), $\mathrm{m_j}$ is the total number of categories (e.g., *strongly disagree, disagree, agree, strongly agree*), and $k = 1, \ldots, m_j$. The item discrimination (α_j) is the set to be the same across the thresholds (τ_h) in an item but α_{j} can vary across items.

Graded Response. Instead of finding the probability of a category response score x_j, the Graded Response (GR) model specifies at the probability of a category score x_j or higher **(**Samejima, 1969). The GR model has been used with rating scale items where the higher categories are associated with more of the latent trait. The probability of a score of x_j or higher can expressed (de Ayala, 2009):

$$p_{x_j}^{*}(\theta) = \frac{e^{\alpha_j\left(\theta-\delta_{x_j}\right)}}{1+e^{\alpha_j\left(\theta-\delta_{x_j}\right)}}, \qquad 2.25$$

where θ is the latent trait, α_j is the item discrimination (varies by item *j*), δ_{x_j} is the category boundary location and need to be in an increasing order. The GR, GPC, and PC are helpful models for polytomous items but limited to unidimensional latent trait (θ).

Multidimensional IRT. Multidimensional IRT (MIRT) models use more than one ability or trait (θ) in the model (e.g., has more than one dimension). There are two types of MIRT models: compensatory that allows the dimensions to interact (e.g., in a survey setting, one trait can be compensated for an abundance of another trait) or non-compensatory (e.g., in a survey setting, one trait will not compensate for another trait). Both types of MIRT models give "an approximation to the relationship between persons' capabilities and the responses to [survey] items" (pg. 58, Reckase, 2009) and Reckase (2009) also notes that there non-compensatory models are best to be called partially compensatory models. These MIRT models are of interest for future studies for the social emotional measures where the skills and supports are multidimensional

and are hierarchical as well. The compensatory model has been applied to research studies (Reckase, 2009) and will be used for the MIRT models in this study.

CM2PL. The compensatory multidimensional two parameter logistic (CM2PL) model is an extension of a 2PL with more than one latent trait and defines a multidimensional item response surface (McKinley & Reckase, 1983; Reckase, 1985). The CM2PL can be written as (de Ayala, 2009):

$$\mathrm{p}\left(\mathrm{x}_{ij}=1\middle|,\underline{\theta}_i,\underline{\alpha}_j,\gamma_j\right)=\frac{\mathrm{e}^{(\underline{\acute{\alpha}}_j\underline{\theta}_i+\gamma_\mathrm{j})}}{1+\mathrm{e}^{(\underline{\acute{\alpha}}_j\underline{\theta}_i+\gamma_\mathrm{j})}} \qquad 2.26$$

where $\mathrm{p}\left(\mathrm{x}_{ij}=1\middle|,\underline{\theta}_i,\underline{\alpha}_j,\gamma_j\right)$ is the probability of a response of 1 (e.g., correct response) to item *j* by person *i*, $\underline{\theta}_i$ is the column vector of person location on each of the F-dimensions, $\underline{\acute{\alpha}}_j$ is a row vector of discrimination parameters for item *j*, and γ_j is the item intercepts. The item intercept (γ_j) can be expressed as (McKinley & Reckase, 1983):

$$\gamma_\mathrm{j}=-\sum_{\mathrm{f}=1}^{\mathrm{F}}\alpha_{jf}\delta_{jf}\,, \qquad 2.27$$

where α_{jf} is the discrimination parameter for item *j* on dimension *f*, δ_{jf} is the difficulty parameter for item *j* on dimension *f*, F is the total number of dimensions being modeled (e.g., empathy, respect for others, appreciation of diversity, organizational awareness). The item intercept is the interaction between the item location (δ_{jf}) and item discrimination (α_{jf}).

MPC. The multidimensional partial credit (MPC) model is for polytomous items that have more than one response category. The compensatory multidimensional two parameter partial credit model (M-2PPC) is for dichotomous items described by Yao and Schwarz (2006). For rating scale items the MPC can be represented by (Reckase, 2009):

$$\mathrm{p}\left(x_{ij}=k\middle|\theta,\delta_{jfk}\right)=\frac{\mathrm{e}^{\sum_{\mathrm{f}=1}^{F}(\theta_{if}-\delta_{jfk})}}{1+\mathrm{e}^{\sum_{\mathrm{k}=1}^{K_j}(\theta_{if}-\delta_{jfk})}}, \tag{2.28}$$

where the where $\mathrm{p}\left(\mathrm{x}_{ij}=k\middle|\theta,\delta_{jfk}\right)$ is the probability of providing a response in item *j* with category k*th*, θ_{if} is the person ability on dimension f, δ_{jfk} is the item difficulty on dimension *f* for score category *k*, K_j is the maximum score for item *j*, *k* = 1, 2,.., K_i. The items vary in difficulty for different dimensions but the item difficulty is the same for response categories within a dimension.

Bifactor. Rijmen (2010) showed a bifactor model in the IRT framework accounts for the variation in participants responses to a set of items in a coordinate space by a general dimension (θ_0) and one of the specific domain dimensions (θ_s). In other words, θ_0 is an estimate of the participant's ability level on the general dimension and θ_s is an estimate of the participant's ability level on the specific domain dimension. A bifactor extension of the full multidimensional GPC model can be written as (Cai, Yang, & Hansen (2011):

$$\mathrm{p}\left(\mathrm{x}_{jk}=k\middle|\theta_{\mathrm{i}0},\theta_{is},\alpha_{j0},\alpha_{js},\gamma_k\right)=\frac{\mathrm{e}^{(\alpha_{j0}\theta_{i0}+\alpha_{js}\theta_{is}+\gamma_k)}}{\sum_{l=0}^{K-1}\mathrm{e}^{(\alpha_{j0}\theta_{j0}+\alpha_{js}\theta_{js}+\gamma_l)}} \tag{2.29}$$

where $\mathrm{p}(x_{jk}|\theta_{\mathrm{i0}}, \theta_{is}, \alpha_{j0}, \alpha_{js}, \gamma_k)$ is the conditional response probability of providing a response in item *j* with category k*th*, θ_{j0} is the person ability on the general dimension, θ_{js} is the person ability on the specific domain dimension, α_{j0} is the item slope to the general dimension, α_{js} is the item slope to the specific domain dimension, and γ_k is the category intercept. When the slope parameters are α_{j0}=1 and $\alpha_{js} = 1$ the items are equally discriminating on the general and specific latent dimensions and becomes the partial credit model for bifactor IRT.

Parameter estimation. There are several estimation methods used for the Rasch model. Most software programs use the maximum likelihood methods, so a few estimation methods using the maximum likelihood methods are briefly presented. The joint maximum likelihood estimation (JMLE) estimates both the person (*i* = 1, …, I) and item (*j* = 1, …, *J*) parameters simultaneously (de Ayala, 2009). However, in JMLE, estimates for perfect scores (e.g., all correct responses) or zero scores (e.g., all incorrect responses) are not used to estimate the item locations. The marginal maximum likelihood estimation (MMLE) attempts to estimate the item parameters and not the person parameters (de Ayala, 2009). The item parameters are estimated using maximum likelihood estimation (MLE) or a Bayesian approach (maximum a posteriori, MAP; expected a posteriori, EAP; de Ayala, 2009). Unlike JMLE, the EAP location estimates all estimated for response patterns of perfect and zero scores. Conditional maximum likelihood estimation (CMLE) method assumes a distribution of the latent variable to estimate the item parameter estimation (van der Linden, 2017).

Model fit. The goodness of fit indices "summarize the discrepancy between the values observed in the data and the values expected under a statistical model" (Reise & Revicki, 2015, p. 112). There are several statistics to evaluate model fit. One is the item characteristic curve (ICC) for dichotomous items that is measured by the chi-square goodness-of-fit index and option response function (ORF) for polytomous items. The likelihood ratio statistic works well for items with categories of four or less. The assessment of item fit uses indices such as Pearson squared statistics that identify poorly fitting items. The global measures of model-data fit helps to compare the fit of two or more different models. The assessment of person-fit helps identifying examinees whose response patterns of item scores are different than the expected response pattern scores.

Dimensionality. Edwards et al. (2015) in chapter 8 from *The Handbook of IRT Modeling*, made a distinction between conceptual dimensionality and empirical dimensionality. Conceptual dimensionality is how the research conceptualizes the operation of the construct and there may be many competing conceptual models with different dimensional structures. Empirical conceptuality refers to the observed statistical properties when a psychometric model is applied to the data. This is where local independence (similar to CFA) applies where once the latent trait is partialled out from the items, then the items are uncorrelated with each other. Paraphrasing Edwards et al., "local independence implies that the only reason items correlate with one another is their association(s) to the common factor(s)." Additionally, conceptual dimensionality is based on the framework the researcher chooses for the construct.

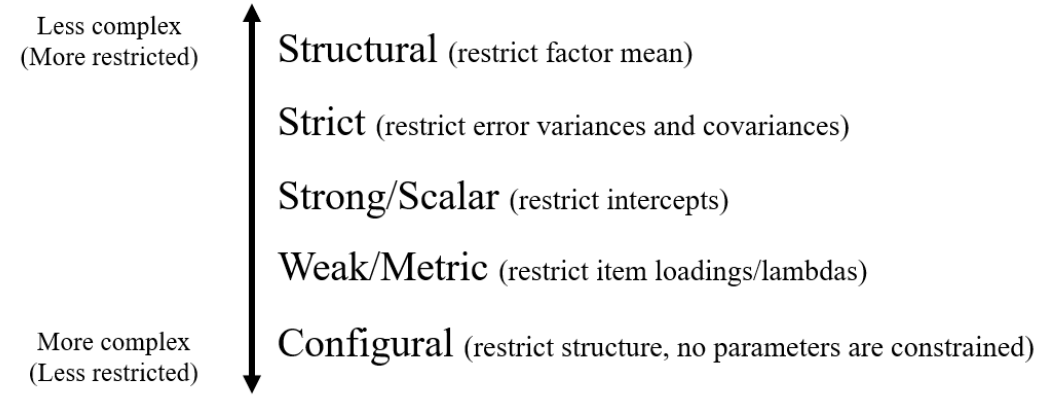
Empirical dimensionality consists of plausible models that agree with the conceptual dimensionality.

Measurement Invariance (MI)

MI is a form of validity evidence that states scores from the quantitatively operationalized SEL construct have the same meaning under different group memberships. Group memberships in this study are the eight ethnic/race groups. There are five levels of MI (from more complex to less complex models): Configural, weak/metric, strong/scalar, strict, structural invariance. Of the five levels of MI, the configural invariance is the least restrictive (most complex) model and structural invariance is the most restrictive (less complex; see Figure 2.8).

Figure 2.8

Measurement Invariance Levels Ordered from Less Restrictive to More Restrictive



Newsom (2015) explained the different levels of MI using the CFA framework. In configural MI the comparison of the variance-covariance matrices are compared across groups to see if the structure of the model appropriate in each of the groups. However, in configural MI level, the loadings, intercepts, error variances and covariances, and factor mean are free to vary. Next, in weak MI the loadings

(unstandardized pattern coefficients) are constrained for each LRV, but the intercepts, error variances and covariances, and factor mean are free to vary. In strong MI, the intercepts (or thresholds) and loadings are constrained and the error variance/covariances and mean factor(s) is/are free to vary. In strict MI the loadings, intercepts, and measurement residuals (error variances and covariances) are restricted but the factor mean(s) is/are freely estimated. The most restrictive MI is structural MI that involves restricting the factor mean(s), error variances/covariances, intercepts, and loadings.

The likelihood ratio test is used to compare the nested models where model M0 is nested within M1 (less restricted model). The null hypothesis states there is no difference in fit between M0 and M1 and the alternative hypothesis states the less restricted model (M1) is an improvement in fit over the more restricted model (M0). For example, for a measurement model to obtain weak MI, the *p-value* needs to be larger than .05 so the data are consistent with the null hypothesis stating there is no difference in fit between the weak MI and configural MI. If, however, the *p-value* is less than .05 then the data are not consistent with the null hypothesis (we reject the null hypothesis). We favor the alternative hypothesis that states the configural MI (less restricted model) is an improvement over weak MI (more restricted model) – the model does not hold weak MI. The desired goal is for the measurement model to obtain strong MI so the scores can be compared across groups.

Cross-sectional Studies

Confirmatory Factor Analysis

CASEL Structure. Ross and Tolan (2018) used the first wave of the National 4-H Study of PYD data to construct a multidimensional CFA using the CASEL (Collaborative for Academic, Social, and Emotional Learning) framework. The CASEL framework has five competencies and the 4-H study included items from five different previously published measures; those items were reorganized into the CASEL domains: Self management (6 items), social awareness (5 items), responsible decision making (9 items), relationship skills (7 items), self-awareness (5 items). The participants were 1,717 students in grade 5 and about 48% males. The race/ethnicity of the sample was composed of African American (8.3%), Asian American (3.6 %), American Indian (4.1%), White (54%), Latino (17.7%), multiracial (5.1%), and other (7.2%).

Before a CFA model was conducted, an explanatory factor analysis was done to see how the items loaded on the five factors. The first CFA model is actually a second-order factor model although the authors called the model a CFA model. The first-order factor structure consists of the five factors and the second-order consists of a general SEL factor. The fit indices are adequate, but not all within the acceptable range ($\chi^2 = 1805, df = 459, p < .01; RMSEA = .041, CFI = .893; TLI = .885$). Since there are 528 elements (32 items, $[32(33 + 1)]/2 = 528$) there were 69 parameters freely estimated ($df = 459$, 528 elements, $528 - 69 = 459$). These parameters include: 27 first-order indicator loadings on the five first-order factors (the

first item in each factor is fixed to 1.00 and there are five factors so five items fixed to 1.00), 32 indicator variances, four second-order loadings to the SEL second-order factor, and six factor variances (five for the first order level and 1 for the SEL second-order factor).

To improve model fit, the model was modified for the seven items of the relationship skills first-order factor to load on two factors and then these two factors were loaded on the relationship skills factor. This would make the second model a third-order CFA model. The first-order factors are a relationship quality factor and create relationship factors that load on the second-order relationship skills factor. However, the rest of the 25 items load on the second-order factors and the third-order factor remains as the SEL factor. Furthermore, some residuals were set to correlate with each other. Model fit indices were good ($\chi^2 = 624, df = 443, p < .001; RMSEA = .015, CFI = .986;\ TLI = .984$). This is the model that was tested for measurement invariance across wave 1, wave 2, and wave 3. However, there is no report of chi-square of difference between model 1 and model 2 for discriminant validity evidence. Also, the software and estimation method of the parameters were not reported. One thing to note is that the CASEL framework is very similar to the Big Five framework (Extraversion, Agreeableness, Conscientiousness, Negative emotionality, Open-mindedness) that comes from personality psychology (Soto & John, 2017).

Higher-order CFA

Five Cs and PYD. Lerner et al. (2005) conducted a longitudinal study, however, only the internal factor structure is reported. The sample of participants are 1,700 fifth graders (47.2% males, mean age = 11.1 years, SD = .53 years; 52.8% females, mean age = 10.9 years, SD = .46 years) from 57 schools and four after school programs. The Five Cs consist of 19 questions with five factors in the first-order level and one factor in the second-order level. The first level has five factors: 1) Confidence, which includes internal sense of having a positive identity and self-worth; 2) Competence, such as having academic and social competence, good grades, being involved in school engagement; 3) Character, that includes personal values, social consciousness, valuing diversity, and interpersonal values and skills; 4) Caring for others and self; and 5) Connection to family, school, community and peers. The factor in the second-order level is PYD. The five factors were nested in the factor of PYD. Note the PYD in this study is a construct as well as a framework. Using ML estimation on raw data within a PRELIS 2.0 file the second-order CFA was conducted in LISREL 8.54 software.

The results of the first model of the second-order latent model had an adequate fit $\chi^2 = 1933, df = 147, p < .01; RMSEA = .085; GFI = .89; CFI = .94; NNFI = .94$. SRMR was not reported in any of the models. Since there are 190 elements (19 items, $[19(19 + 1)]/2 = 190$), there were 43 parameters freely estimated ($df = 147$, 190 elements, $190 - 147 = 43$). These parameters include: 14 first-order indicator loadings to the five factors (the first item in each factor is fixed to 1.00 and

there are 5 factors so 5 items fixed to 1.00), 19 indicator variances, four second-order loadings to the PYD factor (the first loading of the first factor to PYD is fixed to 1.0), and six factor variances (5 for the first order level and 1 for the PYD second-order factor).

Although model 1 had adequate model fit, three other models were tested and ultimately model 3 was chosen. In model 3, the residual errors were allowed to correlate between items within factors and all other estimations were the same as model 1. Brown (2012) commented the correlation of residual errors for items tend to say the correlation between items is explained by something outside the model. The correlation of the residuals seemed to be based after an inspection of the modification indices followed by a brief statement of a theoretical argument. The results for model 3 were a better model fit than model 1, $\chi^2 = 622, df = 136, p < .01; RMSEA = .048; GFI = .98; CFI = .98; NNFI = .98$. Note the df is 136 where there were 11 residual errors estimated.

Big Five. Soto and John (2017) used the Big Five framework to update the Big Five Inventory-2 (BFI-2) that consists of five factors: Extraversion, agreeableness, conscientiousness, negative emotionality that renamed neuroticism, and open mindedness that renamed openness to experience. The participants were 1,000 adults (500 men and 500 women), ranging from ages of 18 to 77, who completed the BFI-2 at the personalitylab.org website. The ethnicity of the sample consisted of White/Caucasian (65%), Hispanic/Latino (7%), Black/African American (7%), Asian/Asian American (6%), Native American/American Indian (1%), another

ethnicity (4%), and multiracial (5%; note with 5% did not response to the ethnicity question). There were 60 rating scaled items, with five categories: 1 (*disagree strongly*), 2 (*disagree a little*), 3 (*neutral; no opinion*), 4 (*agree a little*), and 5 (*agree strongly*).

Each domain had three subdomains referred as facet scales. Each facet scale/ subdomain consisted of 4 items. Extraversion had sociability, assertiveness, energy level (Model 1). Agreeableness consisted of compassion, respectfulness, trust (Model 2). The Conscientiousness facet scales were organization, productiveness, and responsibility (Model 3). Negative emotionality domain had anxiety, depression, and emotional volatility as subdomains (Model 4). Lastly, open mindedness had intellectual curiosity, aesthetic sensitivity, and creative imagination facets (Model 5). After the 15 facets were identified (using principal components analysis), separate results of CFAs were reported where all five three-factor structure (e.g., three fact scales: organization, productiveness, and responsibility) had a better model fit than the respective five one factor structure (e.g., conscientiousness). Also, five bifactor models were reported and had a better fit than the three-factor structure.

The model fit for all five bifactor models are good. Model 1 has $\chi^2 = 332, df = 50, p < .01; RMSEA = .075; BIC = 34556; CFI = .98; TLI = .923.$ Model 2 reports a $\chi^2 = 197, df = 50, p < .01; RMSEA = .054; BIC = 32970; CFI = .952; TLI = .936.$ Next Model 3, $\chi^2 = 338, df = 50, p < .01; RMSEA = .076; BIC = 33428; CFI = .939; TLI = .919.$ Model 4 has $\chi^2 = 328, df = 50, p < .01; RMSEA = .075; BIC = 34377; CFI = .950; TLI = .934.$

Lastly, Model 5 has $\chi^2 = 278, df = 50, p < .01; RMSEA = .068; BIC = 33179; CFI = .930; TLI = .907$.

Extraversion. Instead of using the five constructs from the Big Five framework, Chen et al. (2012) focused on only one construct and its measures. Chen et al. used the bifactor model among other models (e.g., second-order factor model) to illustrate the multifaceted Extraversion construct proposed by Costa and McCrae (1992). Within the Extraversion construct there are six measures or facets called by Chen et. al (2012) that contribute to the one factor structure of Extraversion: Warmth (four items), Gregariousness (four items), Assertiveness (four items), Activity (three items), Excitement Seeking (four items), and Positive Emotions (four items). In the bifactor model, the general factor accounts for all the 23 items underlying the Extraversion factor and the specific domain factors account for the six measures that are the unique variance after partialling out the general factor. The general factor (Extraversion) is orthogonal to the domain specific factors (Warmth, Gregariousness, Assertiveness, Activity, Excitement Seeking, and Positive Emotions) and the specific domain factors are uncorrelated with each other. To scale the factors in the bifactor model, reference indicator was used (e.g., fixing the first factor loading of each factor to 1.00).

Since there are 23 rating scale items there are 253 unique pieces of information, degrees of freedom reported in the study are 209, therefore 67 parameters were estimated (22 item loadings on the general factor, 17 item loadings to the corresponding specific domain factors, 7 factor variances, 21 item residuals and not 23

items residuals). The bifactor model fit adequately, $\chi^2 = 630.09, df = 209, N = 383; RMSEA = .073$ (CI: $.066 - .079$); $SRMR = .058; CFI = .905$. Chen and colleagues noted the Warmth specific domain factor was not significant with a reported variance of .06 and $t = .06$. This suggest Warmth is completely explained by the general factor as there is hardly any variance after extracting the Extraversion general factor.

A second-order factor structure with six first-order factors loaded onto one common factor was applied to the data. The number of estimated parameters were five first-order factors loadings on the common factor (e.g., the first loading of the first-order factors was fixed to 1.00), 17 item loading onto one of the specified first-order factors (e.g., the first item loading in each of the lower order factors was fixed to 1.00), seven item factor variances, 17 item residuals, and six parameters estimated but not clearly mentioned in the study. Model fit was adequate with $\chi^2 = 713.06, df = 224, N = 383; RMSEA = .076$ (CI: $.069 - .082$); $SRMR = .065; CFI = .879$. The bifactor model (more complex, less restricted model) fit better than the second-order model (less complex, more restricted model) with $\Delta\chi^2 = 82.97$ ($\Delta df = 224 - 209 = 15$) , $p < .001$. This suggests the bifactor model may provide a better interpretation of the data than the second-order model.

The participants were 383 undergraduate students. In this study the participants self-identified as 58% women and 42% men; 78.1% European American, 5% African American, 6.5% Asian American, 3.1% Hispanic American, and 7.3% as other. The average age was 19. The researchers cautioned against generalizing the bifactor

internal structure of the Extraversion construct to other populations not reflected in this study.

Chapter III: Methodology

Data Source

This study involves secondary data analysis of a set of items from the 2016 administration of the Minnesota Student Survey (MSS; MN Department of Education, 2018). The MSS is given every three years and the interagency team (MN Departments of Education, Health, Human Services, and Public Safety) designed the MSS to monitor important trends and support planning efforts of the collaborating states agencies, local public school districts, and youth serving agencies and organizations.

Participants

The student survey was anonymously administered to 168,733 public school students in grades 5, 8, 9, and 11. In this study, only grades 8 ($n = 44{,}983$), 9 ($n = 45{,}305$), and 11 ($n = 36{,}576$) were included, because some of the items used in the analysis were not asked of grade 5 students. However, the responses from 126,868 students from grades 8, 9, and 11 were further reduced to 125,959 student responses (explained in the next paragraph) to keep the sample size consistent for measurement invariance analysis. There were 373 students who did not respond to the question, "What is your biological sex?" Of the rest of students, 63,818 students indicated male and 62,677 students indicated female.

The student race and ethnic composition consisted of 4.80% (n = 6,085) American Indian, 3.53% (n = 4,482) Asian Pacific, 4.75% (n = 6,025) Black,

68.47% (*n* = 86,871) White, 9.49% (*n* = 12,040) Latino, 3.60% (*n* = 4,561) multiple race, 2.78% (*n* = 3,522) Hmong, 1.87% (*n* = 2,373) Somali and 0.72% (*n* = 909) race missing. All the analyses (includes CFA, IRT and MI) were conducted without the 909 missing race responses, so the final sample size was 125,959 student responses. This translates to about 35% of responses from grade 8th (*n* = 44,639), 36% of responses from grade 9 (*n* = 44,985), and 29% of responses from grade 11 (*n* = 36,335).

Framework

The Developmental Asset Profile (DAP; The Collaborative for Search Institute, 2005) survey is based on Positive Youth Development (PYD) framework (Benson, 2002). PYD focuses on the strengths of youth developed in their environments. The External Assets in the DAP survey are about the positive relationships youth build with adults and peers in their community. The Internal Assets in the DAP survey are described as competencies, skills, and self-perceptions that are built through different life experiences (Benson, Scales, Hamilton, & Sesma, 2006).

Measures

Internal & External Assets

The DAP survey has 58 items that compose eight developmental assets (four internal, four external). The internal assets are: Commitment-to-Learning (CtL; composed of seven items), Positive Values (composed of eleven items), Social Competencies (composed of eight items), and Positive Identity (composed of six

items). The external assets are: Support (composed of seven items), Empowerment (EM, composed of six items), Boundaries & Expectations (composed of nine items), Constructive-Use-of-time (composed of four items). The MSS uses 31 items of the 58 items from the DAP survey to construct the Developmental Skills and Supports. In addition, the Developmental Skills and Supports have six items that do not overlap with the items that compose the Internal and External Assets. In total, the Developmental Skills and Supports measures are composed of 37 items which are a portion of items given in the MSS.

Developmental Skills and Supports

The Internal and External Assets were reconceived to include Developmental Skills (similar to Internal Assets) and Supports (similar to External Assets) in the MSS items. Developmental Skills and Supports resemble Internal & External Assets but not exactly the same since not all the items in DAP survey were used to compose the Development Skills and Supports measures in the PYD framework. There are three dimensions for Developmental Skills: Commitment-to-Learning (CtL), Positive Identity and Outlook (PI&O), and Social Competence (SC). Developmental Skills (a measure composed of items from the MSS survey) is composed of 20 items and Internal Assets (a measure composed of items from the DAP survey) contain 32 items where these two measures share 14 common items. Developmental Supports (a measure composed of items from the MSS survey) is composed of 17 items and External Assets (a measure composed of items from the DAP survey) has 26 items and these two measures overlap in 17 items. All the 17 items in Development Supports are

the same as External Assets. However, six items in Developmental Skills were not in Internal Assets that prompted into renaming this measure.

CtL, PI&O, SC, EM, FCS and TSS

From the MSS, Developmental Supports can be composed of three dimensions: Empowerment (EM; composed of six items), Family & Community Supports (FCS; composed of five items), and Teacher & School Supports (TSS; composed of six items). Table 3.1 shows the number of items shared between Internal & External Assets measures & Developmental Skills & Supports measures. Notice that the Developmental Skills do not have an equivalence of Positive Values (from Internal Assets) and Developmental Supports does not contain Constructive-use-of-time measure. Although the CtL has the same name in Internal Assets and Developmental Skills, the CtL measure in Developmental Skills is composed of six different items than used in the CtL measure for Internal Assets.

Furthermore, the SEL construct can be conceived of as one general domain (SEL), two domains (Developmental Skills and Supports), or six domains (three Developmental Skills and three Developmental Supports). In addition, the six domains in the Developmental Skills and Supports measure can be conceived of as a single-level (CtL, PI&O, SC, EM, FCS, and TSS**)** or hierarchical (CtL, PI&O, SC within Developmental Skills and EM, FCS, and TSS within Developmental Supports) structure.

Table 3.1

Comparison of Items used in the Internal & External Assets versus Developmental Skills and Supports Measures

Internal/External Asset (number of items)	Developmental Skills/Supports (number of items)	Number of shared items
Internal Assets	**Developmental Skills**	
Commitment-to-learning (CtL; 7 items)	Commitment-to-learning (CtL ; 6 items)	0
Positive Values (11 items)	----------	0
Positive Identity (6 items)	Positive Identity & Outlook (PI&O; 6 items)	6
Social Competencies (8 items)	Social Competence (SC; 8 items)	8
External Assets	**Developmental Supports**	
Empowerment (EM; 6 items)	Empowerment (EM; 6 items)	6
Support (7 items)	Family/Community Supports (FCS; 5 items)	5
Boundaries & Expectations (9 items)	Teacher/School Supports(TSS; 6 items)	6
Constructive Use of Time (4 items)	----------	0

Developmental Supports is about building and being surrounded by positive relationships (Benson, 2006). Within Developmental Supports, FCS and TSS consist of family support, positive family communication, a caring neighborhood, and/or caring school climate. EM is about feeling safe and valued at home, school, and neighborhood. The Developmental Skills are described as competencies, skills, and self-perceptions that are built through different life experiences (Benson, 2006). These include CtL (such as motivation to do well in school, finding school learning useful), PI&O (having a sense of purpose, personal power, positive view of the future) and SC

(planning ahead and decision-making skills, resist negative peer pressure, cultural competence, resolve conflict in a peaceful manner). Developmental Skills are thought to take a longer time to develop than Developmental Supports. Furthermore, Developmental Supports are shaped by the youth's community and Developmental Skills are developed by the individual's personal experiences.

Table 3.2 displays the wording of the 37 items and how they are expected to load on factors. As stated earlier, these factors can be conceptualized as Developmental Skills and Developmental Supports (two factors); or CtL, PI&O, SC, EM, FCS, and TSS (six factors). Most items have a 4-point-rating scale ranging from 1 (*strongly disagree*) to 4 (*strongly agree*) and high scores on the scale indicates the student endorsing the item. A few items have a 5-point-rating scale where 1 = *not at all*, 2 = *a little*, 3 = *some*, 4 = *quite a bit*, and 5 = *very much*. Items Y19 and Y21a had originally four categories, however, category 4 had too few respondents in several ethnic/race community groups. For the measurement invariance analysis, category 4 was merged with category 3 for these two items (both pertain to the CtL measure). Items Y19 and Y21a were left with three categories for all other analyses for consistency. Item Y8r and Y20r were reversed coded so higher scores indicate endorsing the item. There are between five to eight items that load on to the six domains of the Developmental Skills and Supports factors. For the analysis the categorical weighted least squares estimation (weighted least square mean and variance (WLSMV) in Mplus 8.3) method is used since the items are ordered categorical indicators (Rhemtulla, Brosseau-Liard, & Savalei, 2012).

Table 3.2

Wording of the 37 Items and Expected Loading of Items onto Factors

Item	Two factor	Six factor	Item content
Y18	SK	CtL	How often do you care about doing well in school?
Y19	SK	CtL	How often do you pay attention in class?
Y20r	SK	CtL	How often do you go to class unprepared?
Y21a	SK	CtL	If something interests me, I try to learn more about it.
Y21b	SK	CtL	I think things I learn in school are useful.
Y21c	SK	CtL	Being a student is one of the most important parts of who I am.
Y60a	SK	PI&O	I feel in control of my life and future.
Y60b	SK	PI&O	I feel good about myself.
Y60f	SK	PI&O	I feel good about my future.
Y60g	SK	PI&O	I deal with disappointment without getting too upset.
Y60h	SK	PI&O	I find good ways to deal with things that are hard in my life.
Y60n	SK	PI&O	I am thinking about what my purpose is in life.
Y60c	SK	SC	I say no to things that are dangerous or unhealthy.
Y60d	SK	SC	I build friendships with other people.
Y60e	SK	SC	I express my feelings in proper ways.
Y60i	SK	SC	I plan ahead and make good choices.
Y60j	SK	SC	I stay away from bad influences.
Y60k	SK	SC	I resolve conflicts without anyone getting hurt.
Y60m	SK	SC	I accept people who are different from me.
Y60q	SK	SC	I am sensitive to the needs and feelings of others.
Y22b	SP	EM	I feel safe at school.
Y22c	SP	EM	I feel safe in my neighborhood.
Y22d	SP	EM	I feel safe at home.
Y60l	SP	EM	I feel valued and appreciated by others.
Y60o	SP	EM	I am included in family tasks and decisions.
Y60p	SP	EM	I am given useful roles and responsibilities.
Y8r	SP	FCS	Talk with mom.
Y59a	SP	FCS	How much do you feel your parents care about you?
Y59b	SP	FCS	How much do you feel other adult relatives care about you?
Y59c	SP	FCS	How much do you feel friends care about you?
Y59e	SP	FCS	How much do you feel adults in your community care about you?
Y21d	SP	TSS	Overall, adults at my school treat students fairly.
Y21e	SP	TSS	Adults at my school listen to the students.
Y21f	SP	TSS	The school rules are fair.
Y21g	SP	TSS	At my school, teachers care about students.
Y21h	SP	TSS	Most teachers at my school are interested in me as a person.
Y59d	SP	TSS	How much do you feel teachers/other adults at school care about you?

Note. Two factor model: Developmental Skills (SK) or Developmental Supports (SP). Six factor model: Commitment-to-learning (CtL), Positive Identity and Outlook (PI&O), Social Competence (SC), Empowerment (EM), Family/Community Supports (FCS) and Teacher/School Supports (TSS).

Grades

Self-reported grades of students were used to indicate academic performance typically achieved in school. This is on the traditional 4-point scale. The original grade item had the rating values of 1 and 2 be associate with higher grades, however, this item was reversed coded to match the theme of higher scores associate with good grades. Reverse coding of the rating item (Y14r) meant the lower end of the rating scale consistent of *Mostly Incompletes/Mostly Fs* and the higher end of the scale consists of *Mostly Bs/Mostly As*. High scores on the grade rating scale indicates a letter grade of an A. The Grades item is used as a criterion variable.

Race

The students self-reported their ethnic-race background. The ethnic-race community groups are American Indian, Asian Pacific, Black, White, Multiple Races, Latino, Hmong, and Somali. The MSS is unique where there are distinctions between the Asian Pacific students and Hmong students as well as African American (Black) students and Somali students. However, all Latino students are combined, even though Latino students represent different parts of Mexico, Central and South America (Knight, Roosa, Calderon-Tena & Gonzales, 2009). These community groups are mutually exclusive where every student can belong to only one community group.

Scaling Method

The LRVs have non-linear relationships with the indicators (polytomous items) but linear relationships with the factors. The scaling of LRVs used conditional parameterization (theta scaling) where the variance of the LRVs was fixed to 1.00 and the residual variance was freely estimated. The factors in the CFA models were scaled using the unit-loading identification (reference variable) method. The first loading in each factor was fixed to 1.00 and the remainder of the pattern coefficients in the same factor are freely estimated. The reference variable method also allows for the estimation of the factor variances.

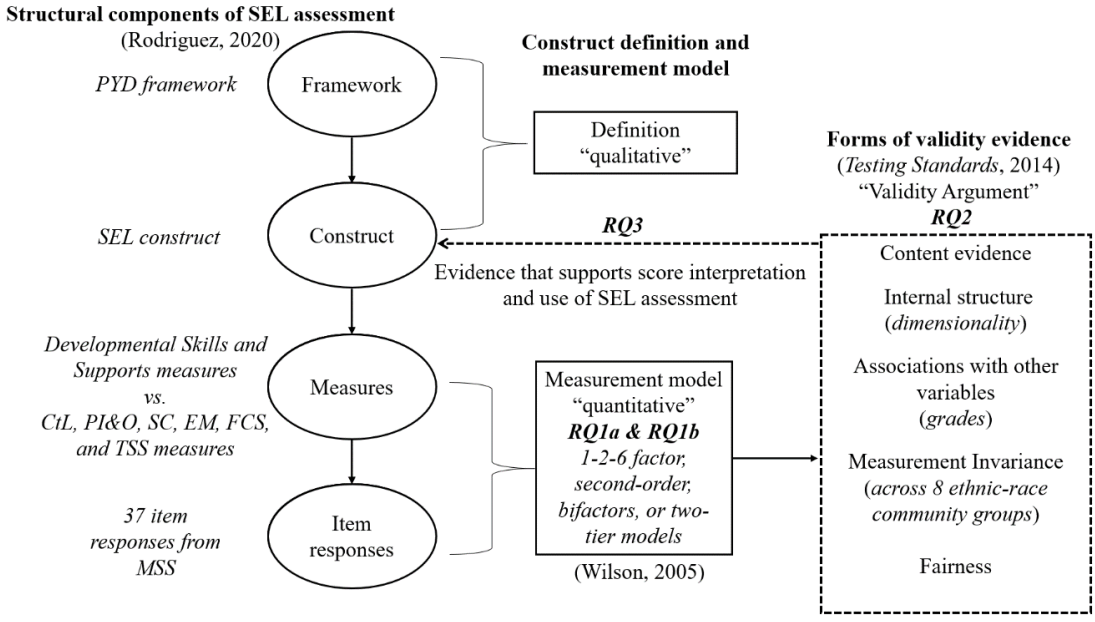
Analytic Models

This study is essentially a validity study for SEL constructs. The validity study proposes a paradigm introduced in chapter 1 (Figure 1.1). It starts at the framework level which provides a qualitative structure of the SEL construct – PYD framework (the first top oval in Figure 3.1). The multifaceted SEL construct (the second oval from the top in Figure 3.1) is operationalized through one measure (general SEL), two measures (Developmental Skills and Supports) or six measures (CtL, PI&O, SC, EM, FCS, and TSS) seen in the third oval from the top in Figure 3.1. However, these measures cannot be directly observed but indirectly observed by the 37 rating scale item responses (the last and bottom oval in Figure 3.1). The participant (in this case a youth) is assumed to have some amount of the SEL construct aligned with PYD that causes the responses to the items and these responses are scored and mapped out to

infer about underlying continuum of the SEL construct or proportion of the SEL construct following Wilson (2005) measurement model. Also, in the measurement model, the scoring of the responses is done that provides plausible quantitative internal structures that are consistent with the PYD framework (the qualitative structure of the SEL construct).

Figure 3.1

An Application of the Paradigm for SEL Assessment



The forms of validity (a unitary concept that can include content evidence, internal structure, associations to other variables, measurement invariance, and fairness displayed in the dotted box in Figure 3.1) are collected at the measure level and inferred at the construct level. The internal factor structure of the measures investigated are: 1-2-6 factor model, second-order model, bifactor models, and two-tier model. The best fitting model is further investigated for MI across eight ethnic-race community groups. The association with grades is briefly investigated with the

best fitting model. The forms of validity are used as evidence to support the score interpretation and use of the SEL construct (e.g., IAU argument). Since there is an established framework of the multifaceted SEL construct in PYD, multidimensional and hierarchical CFA and restricted IRT models are proposed. In other words, the proposed psychometric models are plausible competing measurement models that agree with the qualitative structure of the SEL construct. These psychometric models are presented under each of the three research questions.

Research question 1a

Using the SEL construct based on the PYD framework, there can be one, two or six domains for the SEL construct. Therefore, the first part of research question 1 consists of the plausible competing CFA models for the SEL internal structure (a form of validity evidence). The proposed models are presented from least complex (1 factor model) to most complex (two-tier model). The one factor model represents a unidimensional SEL construct. The correlated traits model can be used to represent a two factor (Developmental Skills and Supports) and six factor (CtL, PI&O, SC, EM, FCS, TSS) model. Model fit, item loadings onto the SEL domain(s), and correlations among the domains are reported and the association with Grades is also briefly inspected.

The second-order model and bifactor model are of hierarchical structure. In the second-order model, the 37 items load onto one of the six first level factors (CtL,

PI&O, SC, EM, FCS, TSS) and the six first-level factors load on the common factor (SEL). The second-order equation can be expressed as (Chen, West & Sousa, 2006):

$$\boldsymbol{\eta} = \boldsymbol{\Gamma\xi} + \boldsymbol{\zeta} \quad \text{eq. 1}$$

$$\boldsymbol{Y} = \boldsymbol{\Lambda_y\eta} + \boldsymbol{\epsilon} \quad \text{eq. 2}$$

Equation 1 is for the common factor structure (Developmental Skills and Supports) represented by $\boldsymbol{\xi}$, vector $\boldsymbol{\eta}$ contains the six lower-order factors (CtL, PI&O, SC, EM, FCS, TSS), the $\boldsymbol{\Gamma}$ vector contains the lower-order factor loadings on the common factor and the $\boldsymbol{\zeta}$ vector represents the disturbances of the lower-order factors (the variance not accounted by the common factor. Equation 2 is for the measurement model for the observed variables that contain six lower-order factors represented by $\boldsymbol{\eta}$, $\boldsymbol{\Lambda_y}$ contains the item loadings onto one of the six lower-order factors and $\boldsymbol{\epsilon}$ vector represents the residuals corresponding to the items.

In the bifactor model, all the 37 items load onto the general factor (SEL) and each item also loads onto one of the six intended domain specific factors (CtL, PI&O, SC, EM, FCS, TSS). The general factor and domain specific factors are orthogonal, which brings-up the challenge of interpretation of the domain specific factors (Rodriguez, M., personal communication, December 20, 2019). The bifactor equation can be expressed as (Chen et al., 2006):

$$Y = \Lambda_y\boldsymbol{\eta} + \boldsymbol{\epsilon},$$

where vector *Y* contains the 37 indicator variables; matrix Λ_y contains the factor loadings of the general and domain specific factors; vector $\boldsymbol{\eta}$ contains the general and domain specific factors; and vector $\boldsymbol{\epsilon}$ contains the residual variance

(includes measurement error and variance not explained by the general factor and domain specific factors). Bifactor model 1 consists of two domain specific factors where Λ_y has the factor loadings of the general factor (SEL), two domain specific factors (Developmental Skills and Supports), vector $\boldsymbol{\eta}$ has the general and two domain specific latent factors, and $\boldsymbol{\epsilon}$ has the residual variance.

Bifactor model 2 consists of six domain specific factors where Λ_y has the factor loadings of the general factor (SEL), six domain specific factors (SC, PI&O, CtL, TSS, FCS, and EM), vector $\boldsymbol{\eta}$ has the general and six domain specific latent factors, and $\boldsymbol{\epsilon}$ has the residual variance.

The two-tier model is also a hierarchical model. First, the items load onto one of the two primary factors (Developmental Skills or Supports) and each item also loads onto one of the six domain specific factors (CtL, PI&O, SC, EM, FCS, TSS). The two primary factors are correlated; however, the primary factors and domain-specific factors are orthogonal to each other. Also, the domain-specific factors are mutually orthogonal to each other. The standard two-tier factor equation can be expressed (Reise & Revicki, 2015)

$$Y = \boldsymbol{\Lambda_y \eta} + \boldsymbol{\epsilon},$$

where vector *Y* contains the 37 indicator variables; matrix $\boldsymbol{\Lambda_y}$ contains the factor loadings of the two primary factors (Developmental Skills or Supports) and domain specific factors; vector $\boldsymbol{\eta}$ contains the primary factors ($\boldsymbol{\eta}$**)** in the first-tier and domain specific factors (CtL, PI&O, SC, EM, FCS, TSS) in the second-tier; and $\boldsymbol{\epsilon}$ contains the residuals.

For the two-tier model, the Metropolis-Hastings Robbins-Monro (MH-RM) estimation algorithm was applied in *FlexMIRT 3.5*. The other CFA models (including hierarchical models) the WLSMV estimator in *Mplus 8.3* (Muthén & Muthén, 2019) was used since the items are ordered categorical indicators (Rhemtulla, Brosseau-Liard, & Savalei, 2012). In all three hierarchical models the residuals associated with the items are uncorrelated.

Research question 1b

The IRT models include unidimensional and multidimensional models for polytomous items. These IRT models are similar in structure to the two-factor, six-factor, and bifactor models that may help in accommodating survey item polytomous scoring. Since the item-response construct map is needed to find the underlying continuum(s) of the construct, the scoring of the items constrained to 1.00 for item discriminations. The unidimensional and multidimensional partial credit model assesses the interaction between item and person parameter – the ordering of persons and items. Both research question 1a (RQ1a) and research question 1b (RQ1b) correspond to the measurement model depicted by a rectangle in Figure 3.1.

Research question 2

The second research question addresses MI that is a form of validity evidence depicted by the dotted square box in Figure 3.1 as RQ2. The proposed SEL construct will be used to compare racial/ethnic community groups so the internal structure of the SEL construct needs to have evidence of strong/scalar level of MI. First, for the

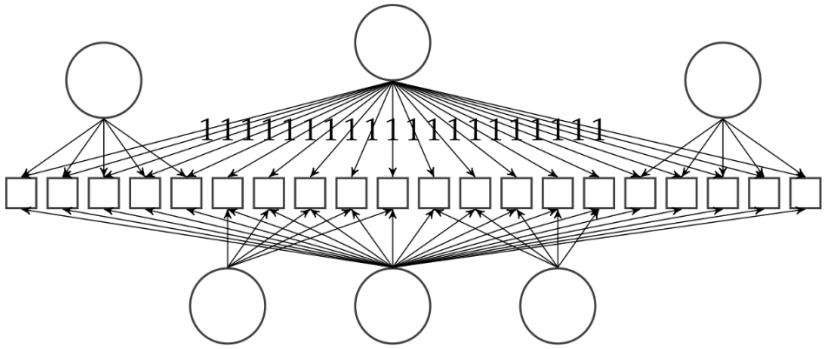
proposed model to achieve configural MI indicates the internal structure of the SEL measure is the same dimensional structure across racial/ethnic groups. Meeting weak/metric MI shows that the strength of associations between items factors and latent response variables are consistent across racial/ethnic groups. The goal is for the proposed measure to obtain strong/scalar MI (the added constraint to weak invariance that the unstandardized intercepts are the same across racial/ethnic groups). Strong/scalar invariance indicates that two people from two different racial/ethnic groups (e.g., White vs. African American) with the same score on a factor (e.g., CtL) are expected to obtain the same observed score on the indicators (items).

Research question 3a

If MI invariance is not met by the best fitting model, then a two-tier model is used to find the variance due to the magnitude of the effect of individual idiosyncratic response style. Figure 3.2 depicts a two-tier model with a random intercept that is orthogonal to the primary factor (e.g., SEL) and domain specific factors (e.g., CtL, PI&O, SC, EM, FCS, TSS). This model requires careful interpretation of results since the variance of the random intercept (effect of individual idiosyncratic response styles) is orthogonal to the primary and specific domain factors. If the variance of the intercept is large, then the lack of MI may be due to response styles. We then need to ask: How will SEL scores be interpreted and used knowing that unstandardized coefficients (patterns and intercepts) are not the same across racial/ethnic groups? If the variance of the intercept is small, the lack of MI may not be due to response styles and this may lead to a bigger problem – the items/response options are not interpreted

the same across groups. Perhaps students in different ethnic community groups score higher than other across different domains.

Figure 3.2
Random Intercept Item Bifactor Model as a Two-tier Model



Note. Adapted from Cai (2010).

Research question 3b

If MI is not met by the best fitting model and there is no indication of idiosyncratic response styles, then maybe the incorrect model was selected (Rodriguez, M., personal communication, February 12, 2020). If this were the case, then the next-best-fitting model was tested for MI.

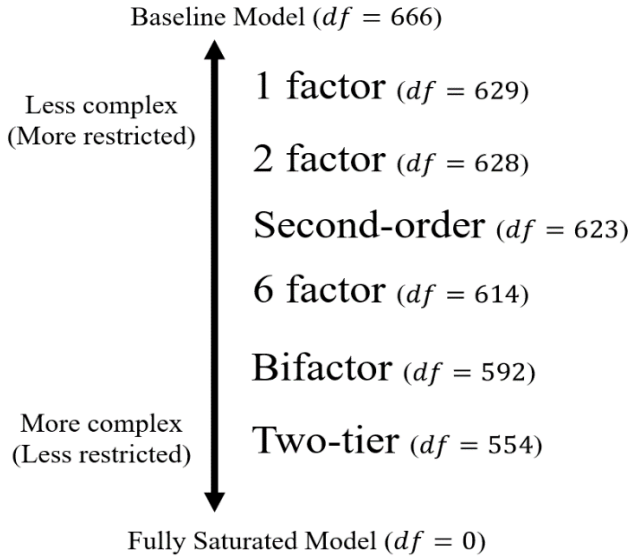
Chapter IV: Results and Analysis

CFA Models

The CFA models is directly tied to research question 1a that aimed to gather pieces of information on the model fit of the internal structure, item loadings of the SEL factors and correlations among the factors. In chapter 1, a paradigm for SEL assessment was proposed as a guide to measure SEL constructs. The PYD framework was used to qualitatively define the SEL construct composed of a measure or measures from item responses of high school students from the MSS. The measurement models were applied at the measures level (e.g., Developmental Skills and Supports) and not at the construct level. All the measurement models are plausible models for score interpretation and use based on the PYD framework. The models are presented from less complex (more restricted) to more complex (less restrictive; see Figure 4.1). Note that in Figure 4.1, the least complex (most restrictive) model is the baseline model with 666 degrees of freedom and the most complex (least restrictive model) is the fully-saturated model with zero degrees of freedom. In this study, the one-factor model is the less complex model and the two-tier model is the more complex model. As stated in chapter 3, the more complex models have more parameters estimated than the less complex models, so the more complex models have fewer degrees of freedoms than the less complex models. Note that the six-factor model has less degrees of freedom (df = 614) than the second-order model (df = 623); therefore, the six-factor model is more complex than the second-order model. Since the bifactor models with two-domain specific factors and six-domain specific factors had the same

number of degrees of freedom (*df* = 592), they are both represented in Figure 4.1 as bifactor.

Figure 4.1
Plausible Models Ordered from Least Complex to Most Complex



1-2-6 Factor models

Due to lack of choosing response option 4 (*strongly agree*), items Y19 and Y21a were recoded from four levels to three levels for all analyses. The trend of the fit-indexes suggests that the model-data fit improves as the number of factors increases (Table 4.1). A six factor model fits better than a two-factor model and a one-factor model.

The less restricted (more complex) model is an improvement over the more restricted (less complex) model when the chi-square difference test is significant (Brown, 2015). The chi-square of difference test between the two-factor model (less restricted model) and one factor model (more restricted model) was significant where the $\Delta\chi^2 = 199512, (\Delta df = 1), p < .001, \alpha = .05$ indicating the two-factor model fit the data better than the one factor model. Furthermore, a chi-square of difference test showed evidence that the six-factor model (less restricted model) was an improvement over the two-factor model (more restricted model), the $\Delta\chi^2 = 334338, (\Delta df = 14), p < .001, \alpha = .05$. All chi-square of difference tests were conducted using the DIFFTEST option in Mplus since the WLSMV estimator was used.

In the one-factor model, there were many covariances not explained by the model. In fact, most of the residual correlations were above .10 that were either large negative or positive residual correlations. This mismatch is not surprising as the one-factor model had a poor model fit. In the two-factor model, most of the larger residual correlations (between .29 and .35) are in Y18 and Y19, Y21c and Y21b, Y60j and Y60c indicators specified to measure the Developmental Skills. The indicators Y22c, Y22d, Y60p, Y60p Y59a, Y59b, Y21e, Y21f, Y21 g, and Y21h tended to have larger residual correlations (between .22 and .37) in the Developmental Supports factor. Specifically, Y22c, Y22d, had the highest residual correlation of .37 Developmental Skills factor. Again, the many high residual correlations were not surprising as the model fit for the two-factor model was poor.

In the six-factor model, there are fewer covariances not explained by the model. For the CtL factor, Y19 and Y18 indicators had a residual correlation of .10. In the PI&O factor, the indicators with residuals over -.10 are: Y60n and Y60b at -.16, Y60h and Y60n at -.11. Indicator 8r had the lowest residual (<.09) with the FCS factor indicators. The indicators for SC had the fewest residual correlations above .10: Y60j and Y60c at .20, Y60j and Y60d at -.16. In contrast, the EM factor had mostly large correlation residuals (between .12 and .28): where 5 (Y22b, Y22c, Y22d, Y60l, Y60p not Y60o) out of 6 items are above .10 and this suggests these indicators are sharing unique variance not related to the EM factor (Kline, 2015) or these items do not demonstrate local dependence. The indicators for factor FCS with moderate to high correlation residuals are Y8r and Y59a at -.12 and Y59e and Y59a at -.17. Lastly, the indicators that measure TSS have higher correlation residuals (between -.20 and -.22) are: Y59d and Y21d, Y59d and Y21e, and Y59d and Y21f.

Table 4.1

Model Fit Indices for CFA and Higher-Order Models

Fit Measures	CFA Models			Higher-order Models		
	One factor	Two factor	Six factor	Second-order	Bifactor 1	Bifactor 2
χ^2	1128546	929034	594696	666227	752163	474092
df	629	628	614	623	592	592
p-value	<.001	<.001	<.001	<.001	<.001	<.001
CFI	.755	.798	.871	.855	.837	.897
TLI	.741	.786	.860	.846	.816	.884
RMSEA	.126	.115	.093	.098	.106	.084
SRMR	.098	.092	.068	.073	.079	.062

The item factor loadings of the one, two, and six factor models are reported in Table 4.2. All factor loadings in the CFA models were statistically significant where all item loadings are different than zero. In the one-factor model the standardized loadings ranged from .258 (item Y20r) to .802 (item Y60l), the two-factor model standardized loadings were from .274 (item Y20r) to .847 (item Y60l), and these standardized loadings were even higher in the six-factor model which were from .340 (item Y20r) to .942 (item Y59d). In the three CFA models there are two items that had consistently lower loadings than the rest of the items. For the first item, Y20r, the range of standardized loadings is .258 to .340; and for the second, item Y21a, the range of standardized loadings is .376 to .496.

Table 4.2
Standardized Loadings for CFA Models

	One factor	Two factor		Six factor					
Item	SEL	SK	SP	CtL	PI&O	SC	EM	FCS	TSS
Y18	.581	.613		.777					
Y19	.544	.575		.727					
Y20r	.258	.274		.340					
Y21a	.376	.396		.496					
Y21b	.519	.545		.696					
Y21c	.535	.564		.719					
Y60a	.739	.772			.813				
Y60b	.736	.772			.816				
Y60f	.792	.828			.871				
Y60g	.653	.684			.726				
Y60h	.768	.804			.851				
Y60n	.499	.527			.562				
Y60c	.631	.664				.694			
Y60d	.670	.705				.740			
Y60e	.724	.759				.798			
Y60i	.732	.769				.806			
Y60j	.682	.717				.749			
Y60k	.665	.700				.732			
Y60m	.451	.478				.504			
Y60q	.536	.566				.595			
Y22b	.621		.651				.681		
Y22c	.639		.670				.703		
Y22d	.666		.697				.731		
Y60l	.802		.847				.890		
Y60o	.776		.808				.841		
Y60p	.792		.827				.860		
Y8r	.497		.520					.574	
Y59a	.745		.777					.851	
Y59b	.719		.750					.820	
Y59c	.608		.635					.702	
Y59e	.744		.774					.866	
Y21d	.660		.691						.782
Y21e	.689		.721						.810
Y21f	.578		.607						.698
Y21g	.718		.750						.838
Y21h	.655		.684						.776
Y59d	.774		.804						.942

The intercorrelations in the two-factor and six-factor models may give some insight if Developmental Skills and Supports measure is unidimensional (Chang, 2015). In the two-factor model, Developmental Skills and Supports factors are highly associated with each other (disattenuated correlation of .803), which suggests the structure of the Developmental Skills and Supports measure is approaching unidimensional. In the six-factor model, the 12 intercorrelations range between .509 and .841, and three disattenuated correlations are higher than or equal to .80 (see Table 4.3).

Table 4.3
Factor Correlations for Six Factor Model

	CtL	PI&O	SC	EM	FCS
PI&O	.555				
SC	.676	.841			
EM	.602	.802	.806		
FCS	.514	.705	.694	.780	
TSS	.664	.509	.573	.621	.738

Note. Developmental Supports include commitment-to-learning (CtL), Positive Identity and Outlook (PI&O) and Social Competence (SC). Developmental supports variables include Empowerment (EM), Family/Community Supports (FCS) and Teacher/School Supports (TSS).

Second-order model

Each item had a non-zero loading on the first-order factor that it was theorized to measure. The covariance among the first-order factors was explained by the common factor, where lower-order factor loadings ranged from .720 to .901. The second-order model did not have a better model fit than the six-factor model, ($\chi^2 = 666227, df = 623$, p-value < .001, CFI = .855, TLI = .846, RMSEA = .098, SRMR = .073).

Six-factor model (less restricted model) had a better model fit than the second-order factor model (more restricted model). The model fit indices in the second-order factor model were not an improvement over the six-factor model.

Table 4.4

Standardized Loadings for Second-Order Model and First-order Loadings onto Common Factor

	First-order factors					
Item	CtL	PI&O	SC	EM	FCS	TSS
Y18	.779					
Y19	.728					
Y20r	.342					
Y21a	.496					
Y21b	.694					
Y21c	.718					
Y60a		.812				
Y60b		.815				
Y60f		.873				
Y60g		.724				
Y60h		.853				
Y60n		.559				
Y60c			.693			
Y60d			.740			
Y60e			.801			
Y60i			.807			
Y60j			.748			
Y60k			.732			
Y60m			.502			
Y60q			.594			
Y22b				.681		
Y22c				.702		
Y22d				.731		
Y60l				.891		
Y60o				.840		
Y60p				.860		
Y8r					.572	
Y59a					.848	
Y59b					.817	
Y59c					.700	
Y59e					.874	
Y21d						.780
Y21e						.807
Y21f						.695
Y21g						.836
Y21h						.773
Y59d						.953
Common	.720	.860	.895	.901	.856	.717

Bifactor Models

The bifactor model 2 with six specific domain factors ($\chi^2 = 474092, df = 592$, *p*-value < .001, CFI = .897, TLI = .884, RMSEA = .084, SRMR = .062) fits better than the bifactor model 1 with two specific domain ($\chi^2 = 752163, df = 592$, *p*-value < .001, CFI = .837, TLI = . 816, RMSEA = .106, SRMR = .079). Bifactor model 2 had adequate model fit as the CFI and TLI fit indices were close to .90. However, bifactor model 1 had a poor model fit since the CFI and TLI were closer to .80. Bifactor model 1 and 2 have the same degrees of freedom so the χ^2 difference tests for the WLSMV estimator could not be computed (bifactor 1 model was not nested in bifactor 2 model).

The non-zero loadings of the General (G) factor in the bifactor model 2 with six domain specific factors (displayed in Table 4.5) are very similar to the one-factor model (Table 4.2). However, in the bifactor models, some of the loadings on the General factor are smaller than the factor loadings on the items of the specific domains factors. This may indicate that some specific-domain factors explain above and beyond the General factor (bolded items in Table 4.5). In bifactor model 1, there are 6 items on the specific domain factors with negative loadings. However, all the item loadings except item Y18 on the two specific-domain factors were statistically significant.

Since the bifactor model 2 (one general factor and six specific factors) had better model fit indices than bifactor model 1 (one general factor and two specific factors), the bifactor 2 model was chosen to compare with the second-order model.

A chi-square of difference test showed evidence that the bifactor model 2 (less restricted model) is an improvement over the second-order factor model (more restricted model), the $\Delta\chi^2 = 192{,}135\ (\Delta df = 31), p < .001, \alpha = .05.$

Table 4.5

Standardized Loadings for Bifactor Models

	Model 1			Model 2						
	G	Domain Specific		G	Domain Specific					
Item	SEL	SK	SP	SEL	CtL	PI&O	SC	EM	FCS	TSS
Y18	.602	.001		.554	**.579**					
Y19	.563	-.121		.518	**.530**					
Y20r	.270	.037		.254	.185					
Y21a	.387	-.054		.376	.192					
Y21b	.532	-.259		.496	.484					
Y21c	.550	-.233		.512	**.526**					
Y60a	.762	.456		.689		.461				
Y60b	.758	.400		.703		.390				
Y60f	.814	.461		.748		.436				
Y60g	.638	.558		.599		.482				
Y60h	.759	.564		.720		.481				
Y60n	.521	.121		.529		-.042				
Y60c	.634	.282		.579			.515			
Y60d	.694	.236		.697			.034			
Y60e	.748	.376		.746			.103			
Y60i	.755	.310		.734			.259			
Y60j	.688	.277		.624			.611			
Y60k	.688	.263		.647			.380			
Y60m	.471	.091		.432			.361			
Y60q	.558	.081		.525			.325			
Y22b	.511		.472	.609				.455		
Y22c	.522		.497	.579				**.720**		
Y22d	.564		.468	.615				.574		
Y60l	.819		.046	.838				-.034		
Y60o	.806		-.069	.797				-.267		
Y60p	.825		-.083	.815				-.291		
Y8r	.501		.074	.492					.288	
Y59a	.747		.124	.699					.547	
Y59b	.712		.164	.674					.610	
Y59c	.605		.131	.609					.292	
Y59e	.677		.371	.761					.205	
Y21d	.446		**.668**	.508						**.656**
Y21e	.444		**.724**	.509						**.715**
Y21f	.428		**.541**	.477						**.539**
Y21g	.467		**.738**	.541						**.713**
Y21h	.499		**.571**	.563						.540
Y59d	.671		.484	.768						.277

Note: Developmental Skills (SK) and Supports (SP). Commitment-to-learning (CtL), Positive Identity and Outlook (PI&O) and Social Competence (SC), Empowerment (EM), Family/Community Supports (FCS) and Teacher/School Supports (TSS).

The correlation residual matrix was also inspected to assess the local fit of the Bifactor models. In Bifactor model 1 with one general factor and two domain factors, there are many covariances not explained by the model. There were many correlation residuals above .10 for indicators that measure the internal domain factor and few residuals above .10 for the indicators that measure the external domain factor.

In Bifactor model 2 with one general factor and six domain factors, the covariances not explained by the model are:

- Y60a & Y60g, Y60a & Y60h for the PI&O domain factor
- Y60j & Y60m, Y60j & Y60q, Y60m & Y60q for the SC domain
- Y22c & Y60o, Y22c & Y60p, Y22d & Y60o, Y22d & Y60p for the EM domain.

CtL, FCS, and TSS assigned indicators have residual correlations less than .10.

Two-tier Model

The two-tier model with two correlated primary factors and six specific domain factors barely converged using the Metropolis-Hastings Robbings-Monro (MHRM) estimation method in *FlexMIRT* 3.5. The model fit indices reported are the 95% confidence intervals for -2loglikelihood (-2LL), Akaike information criterion (AIC), and Bayesian information criterion (BIC). The -2LL = [9074325, 9074439], AIC = [9074705, 9074819], BIC = [9076556, 9076670]. Both the AIC and BIC penalize for model complexity. Both AIC and BIC relative fit indices are quite high which indicates the two-tier model does not fit the data well. It would have been

desired to obtain lower AIC and BIC values that indicated the two-tier model fits the data well, however, this is far from the current case.

The item loadings are reported in Table 4.6. Three items (Y60c, Y60q, Y22d) had negative standard errors for the domain specific factor loading. Item Y22c had a negative standard error for the primary and domain specific factors. Five items loaded higher on the domain specific factor than the primary factor (items bolded in black; Y21b, Y21c, Y21d, Y21e, Y21f). Most of the loadings in the primary factors were similar to the general factor loadings in the bifactor model 1 (one general factor with two domain specific factors). Overall, most items loaded higher on the primary factors than on the domain specific factors.

Table 4.6

Standardized Loadings for Two-tier with MHRM Estimation

	Primary factors		Domain Specific factors					
Item	SK	SP	CtL	PI&O	SC	EM	FCS	TSS
Y18	.58		.43					
Y19	.63		.47					
Y20r	.21		.07					
Y21a	.40		.21					
Y21b	.46		**.55**					
Y21c	.49		**.58**					
Y60a	.71			.49				
Y60b	.68			.45				
Y60f	.78			.44				
Y60g	.62			.20				
Y60h	.76			.27				
Y60n	.46			-.03				
Y60c	.65				**.**15			
Y60d	.69				-.22			
Y60e	.77				-.15			
Y60i	.80				.05			
Y60j	.97				.25			
Y60k	.73				.03			
Y60m	.49				-.08			
Y60q	.54				**-.**11			
Y22b		.54				.16		
Y22c		**.**93				**.**36		
Y22d		.54				**.**28		
Y60l		.74				.22		
Y60o		.70				.51		
Y60p		.73				.47		
Y8r		.30					.20	
Y59a		1.00					.09	
Y59b		.63					.57	
Y59c		.59					.34	
Y59e		.67					.31	
Y21d		.53						**.64**
Y21e		.54						**.71**
Y21f		.47						**.53**
Y21g		.58						.71
Y21h		.60						.51
Y59d		.67						.21

Note. Developmental Skills (SK) and Supports (SP). Commitment-to-learning (CtL), Positive Identity and Outlook (PI&O) and Social Competence (SC), Empowerment (EM), Family/Community Supports (FCS) and Teacher/School Supports (TSS). Bolded items loaded higher on the domain-specific items than the primary factors.

IRT Results

Research question 1b focused on using IRT models to estimate the item and person location of a latent trait continuum. The IRT models are presented from simple to complex models. The IRT models require complete item response patterns of the measures. The reported samples size is 125,959 participants. However, each measure had missing pattern responses, and these responses were deleted from the final sample size. At the end, a sample of 112,309 responses were used to estimate of item location and person location. The sample size was 89% of the 125,959 sample size. Also, the response rate of African American and Somali ethnic/race community groups were the lowest across all items compared to Native American, Asian Pacific, White, Multiple Races, Latino and Hmong. The item response rate from the Somali group ranged from 77.2% to 98.2% of 1,273 Somali youth. The item response rate from the African American group ranged from 83.7% to 98.6% of 6,025 African American youth. The other six ethnic/race community groups had item response rates between 94.0% to 99.2%.

Unidimensional Partial Credit Models

All the IRT models used the partial credit model. The six measures (CtL, PI&O, SC, EM, FCS, and TSS) were fit to separate IRT models. The model-level fit indices are reported in Table 4.7. From Table 4.7, PI&O and TSS models seem to be better fitting models as they have CFIs and TLIs in the .90s. However, PI&O had the second to highest levels of AIC and BIC values and FCS had the lowest AIC and BIC values (although the AIC and BIC are quite high). PI&O had a low RMSEA of .06

and TSS had the highest RMSEA value of .22.

Table 4.7

Model-fit Statistics for Six Unidimensional Partial Credit Models

	CtL	PI&O	SC	EM	FCS	TSS
LL	-608054	-733202	-934049	-594762	-588457	-630750
EsPa	17	19	25	19	21	20
AIC	1216142	1466442	1868148	1189562	1176957	1261539
AICc	1216142	1466442	1868148	1189562	1176957	1261539
BIC	1216306	1466625	1868389	1189745	1177159	1261732
SABIC	1216252	1466565	1868309	1189685	1177092	1261668
M2	4502	717	22833	7820	6991	56
df	4	2	11	2	14	1
p-value	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
RMSEA	.10	.06	.14	.19	.21	.22
SRMSR	.14	.16	.09	.13	.13	.10
CFI	.63	.96	.44	.88	.79	.99
TLI	.54	.94	.39	.82	.78	.99

Note. EsPa = Estimated number of parameters. Commitment-to-learning (CtL), Positive Identity and Outlook (PI&O), Social Competence (SC), Empowerment (EM), Family/Community Supports (FCS) and Teacher/School Supports (TSS).

To assess item-fit the item location parameters are reported in Table 4.8. Most of the items had locations that suggested that the items were relatively easy to endorse (most item b-parameters and thus thresholds were negative). In other words, the item threshold parameters indicate the relative difficulty of each step (Embretson & Reise, 200). For example, in item Y18, the relative difficulty for the transition between category zero to category one is -3.83 represented by the first threshold 1 (δ_1) in Table 4.8. The transition item parameter indicates the relative difficulty to other transitions within an item.

The option characteristic curves (OCCs) for most of the polytomous items show $\delta_1 < \delta_2 < \delta_3$ for items with three thresholds and $\delta_1 < \delta_2 < \delta_3 < \delta_4$ for items

with four thresholds. Three items did not have ordered thresholds which are Y20r, Y22d, and Y59a.

Table 4.8

Item Location Parameters for Six Unidimensional Partial Credit Models

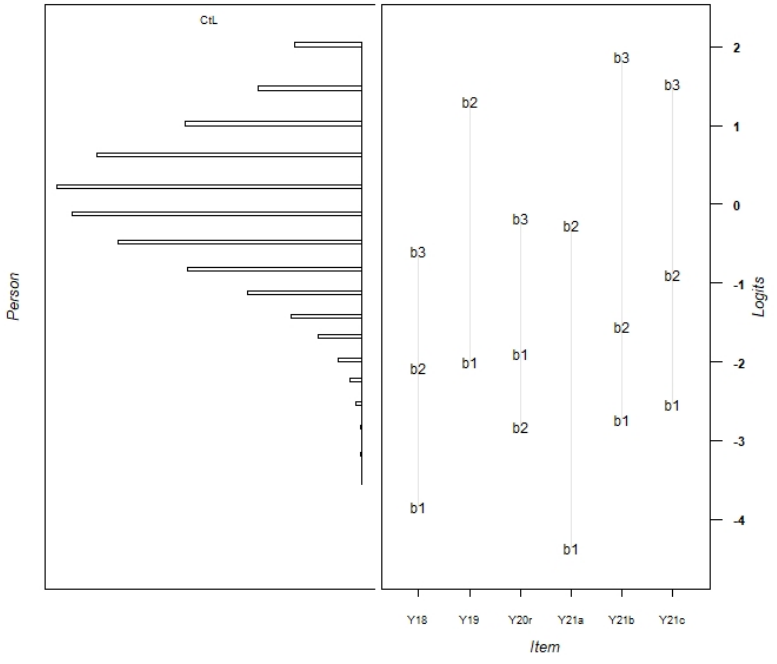
			Item transition locations			
Item	Trait	Label	δ_1	δ_2	δ_3	δ_4
1	CtL	Y18	-3.83	-2.07	-.59	
2		Y19	-2.00	1.31		
3		Y20r	-1.89	-2.82	-0.17	
4		Y21a	-4.35	-0.26		
5		Y21b	-2.72	-1.54	1.87	
6		Y21c	-2.53	-0.88	1.53	
7	PI&O	Y60a	-3.00	-1.17	0.67	
8		Y60b	-2.60	-0.98	0.75	
9		Y60f	-2.90	-1.23	0.52	
10		Y60g	-2.41	-0.44	1.57	
11		Y60h	-2.65	-0.72	1.28	
12		Y60n	-2.66	-1.24	0.42	
13	SC	Y60c	-2.62	-1.14	-0.45	
14		Y60d	-2.83	-1.18	0.49	
15		Y60e	-2.24	-0.57	1.42	
16		Y60i	-2.98	-0.91	0.93	
17		Y60j	-2.63	-1.01	-0.04	
18		Y60k	-2.94	-1.12	0.67	
19		Y60m	-3.51	-2.36	-0.42	
20		Y60q	-2.69	-1.29	0.33	
21	EM	Y22b	-4.04	-3.36	-0.07	
22		Y22c	-4.28	-4.10	-0.73	
23		Y22d	-4.32	-4.39	-1.38	
24		Y60l	-2.86	-0.96	1.10	
25		Y60o	-3.05	-1.46	0.55	
26		Y60p	-3.42	-1.66	0.61	
27	FCS	Y8r	-2.54	-2.49	-1.88	-.82
28		Y59a	-3.04	-2.43	-1.72	-1.76
29		Y59b	-2.90	-2.41	-1.72	-0.70
30		Y59c	-3.00	-2.45	-1.37	0.10
31		Y59e	-1.55	-0.92	0.36	1.35
32	TSS	Y21d	-3.72	-2.33	1.64	
33		Y21e	-4.23	-2.17	2.18	
34		Y21f	-3.95	-1.89	2.25	
35		Y21g	-4.43	-3.03	1.43	
36		Y21h	-3.76	-1.36	2.28	
37		Y59d	-3.15	-1.78	0.09	1.95

Note. Items are grouped by dimension in gray shading or no shading. Items are grouped in sequential order for latent traits.

The item-person map, also called the Wright map, is a useful tool to visually represent the ordering of respondents and items on the same logit scale (Wilson, 2005). The Wright map helps in displaying the underlying continuum of the construct(s). Figure 4.2 is the conceptualization of the unidimensional CtL construct where on the left-hand side are the participants' latent trait levels (labeled as Person) and the right-hand side are the threshold parameters (labeled as b1, b2, and b3) under the Item heading. In general, the range of the items are adequate for the range of the latent trait levels among the participants. Items Y18 and Y21a are located at the lower range of theta. These two items may be easier to endorse on the high level. In other words, items Y18 and Y21a may require lower levels of CtL to endorse at a high level relative to other items. Note the Wright map shows the thresholds for item Y20r are not in ascending order.

Figure 4.2

Item-person Map for the One-dimensional Commitment-to-learning (CtL) Construct



For the PI&O construct displayed in Figure 4.3, the range of items are adequate for the latent trait levels ranging between -3 to 2 logits. The distribution of the latent trait shows there are few students in the trait levels below -3 and above 2 logits.

Figure 4.3

Item-person Map for the One-dimensional Positive Identity and Outlook (PI&O) Construct

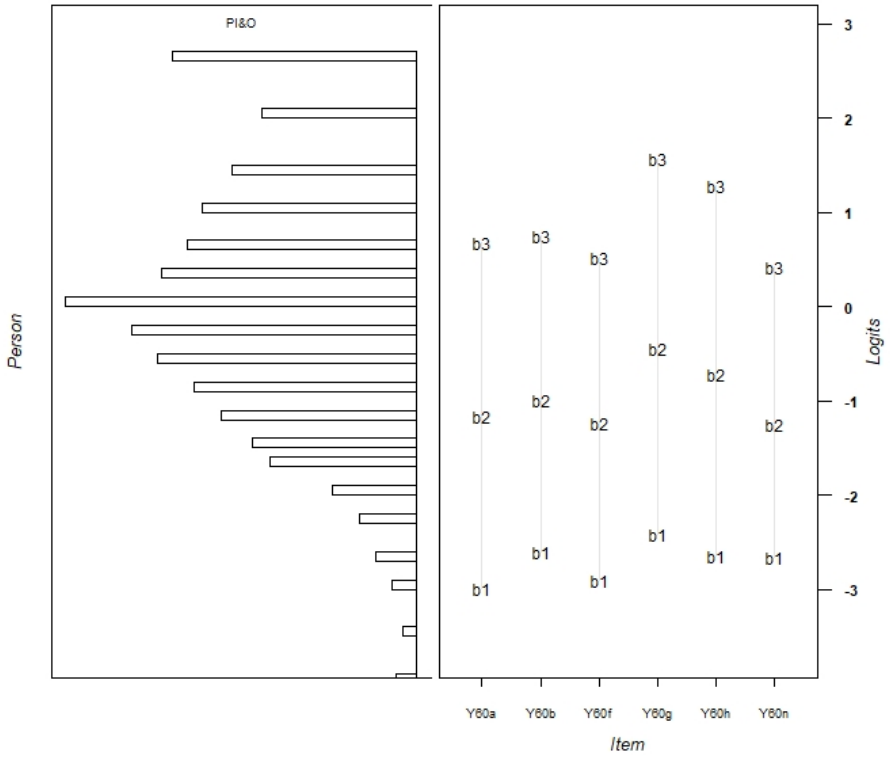


Figure 4.4 displays the underlying continuum for the SC construct. The range of items are adequate for the latent trait levels ranging between -4 to 1.5 logits. The eight items seem to be located at the lower range of theta, so these items may be easier to endorse at a high level. All the items have thresholds in ascending order.

Figure 4.4

Item-person map for the One-Dimensional Social Competence (SC) Construct

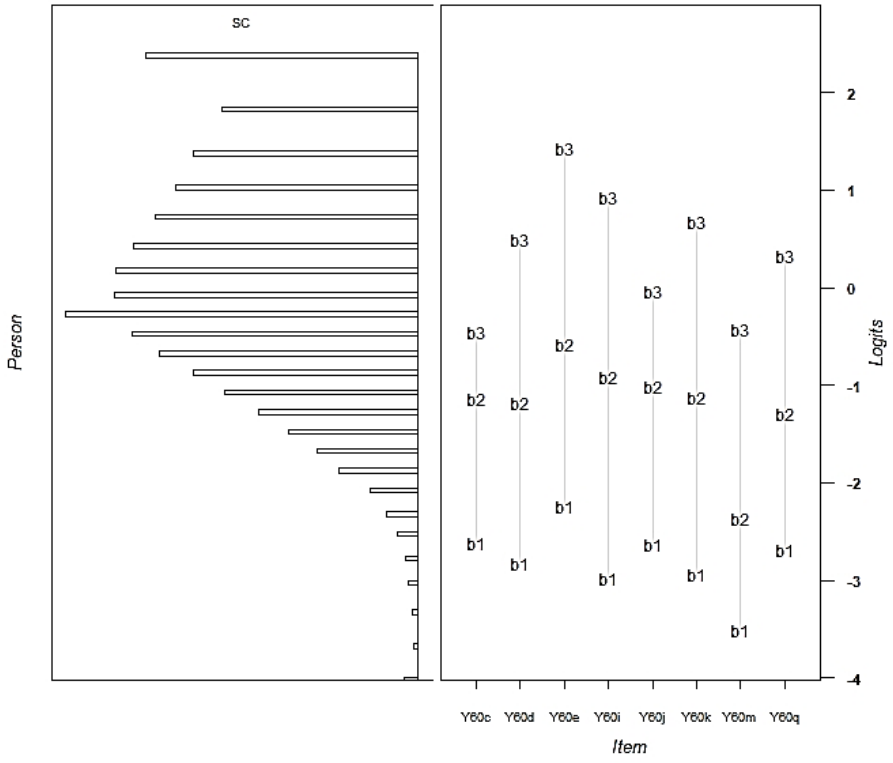


Figure 4.5 displays the underlying continuum for the EM construct. The range of items are adequate for the latent trait levels below 1 logit. More items are needed for latent trait levels above 1.0 logit. All the items have thresholds in ascending order, however, items Y22c and Y22d have thresholds b1 and b2 very close together, reflecting the limited information distinguishing the lower response options.

Figure 4.5

Item-person Map for the One-dimensional Empowerment (EM) Construct

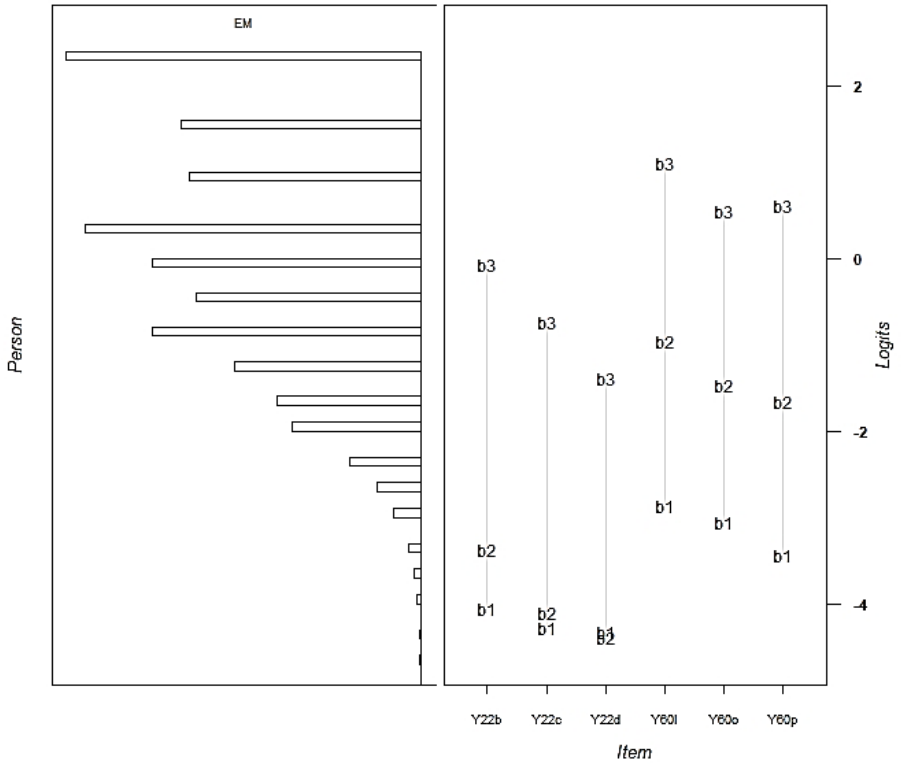


Figure 4.6 displays the underlying continuum for the FCS construct. The range of the items are adequate for the range of the latent trait levels among the participants between -3 and 1.5 logits. There is a larger noticeable distance between score points in the higher end of the continuum for the FCS latent trait variable, reflecting the nonlinear raw to Rasch score transformation. Item 8r has thresholds b1 and b2 that are close together, reflecting the limited information distinguishing the lower response options.

Figure 4.6
Item-person Map for the One-dimensional Family/Community Supports (FCS) Construct

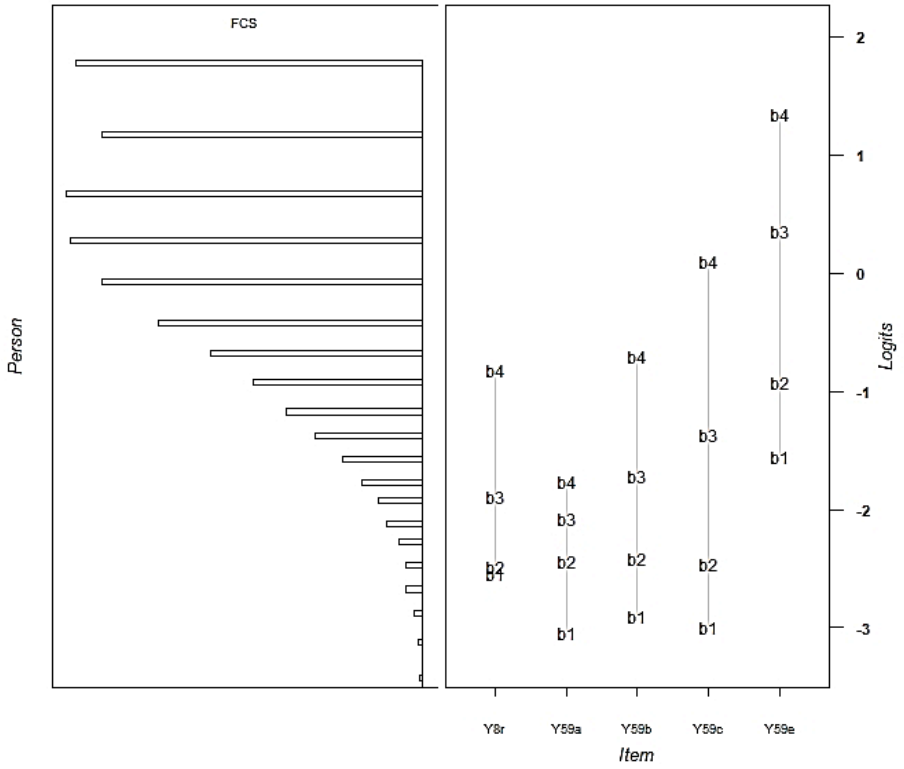
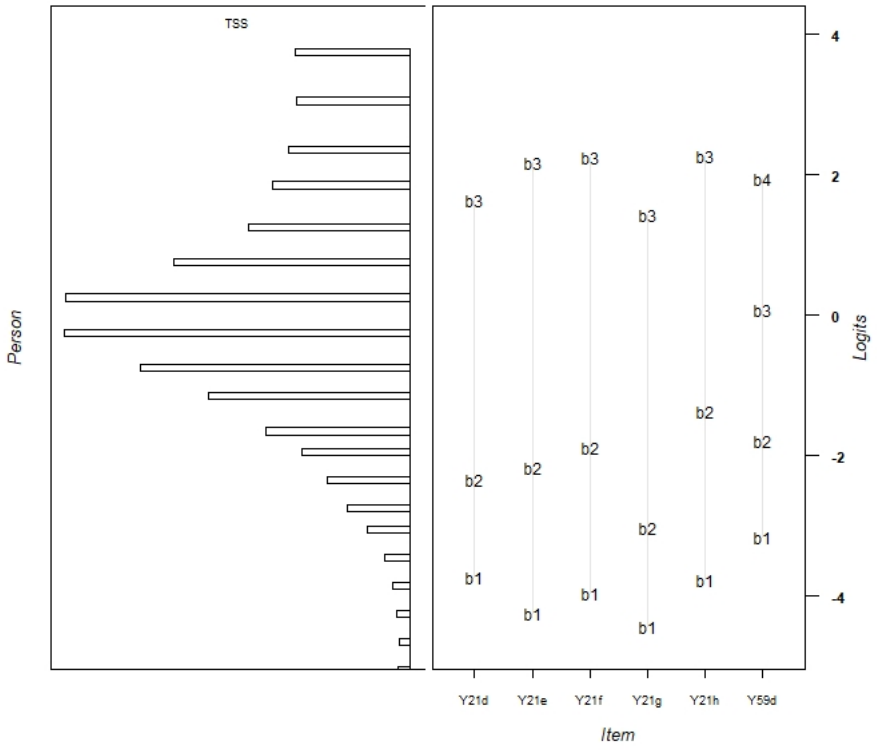


Figure 4.7 displays the underlying continuum for the TSS construct. The range of the items are adequate for the range of the latent trait levels among the participants between -4.5 and 2 logits. All the item thresholds are ordered.

Figure 4.7

Item-person Map for the One-dimensional Teacher/School Supports (TSS) Construct



Next, a unidimensional IRT model included all the items in one latent trait. The item transition locations are reported in Table 4.9. Compared to the item transition locations in Table 4.8, all but four items had ordered thresholds (e.g., $\delta_1 < \delta_2 < \delta_3$). Items Y20r, Y22c, Y22d, and Y8r did not have ordered thresholds, and note that Y20r and Y22d did not have ordered thresholds in Table 4.8.

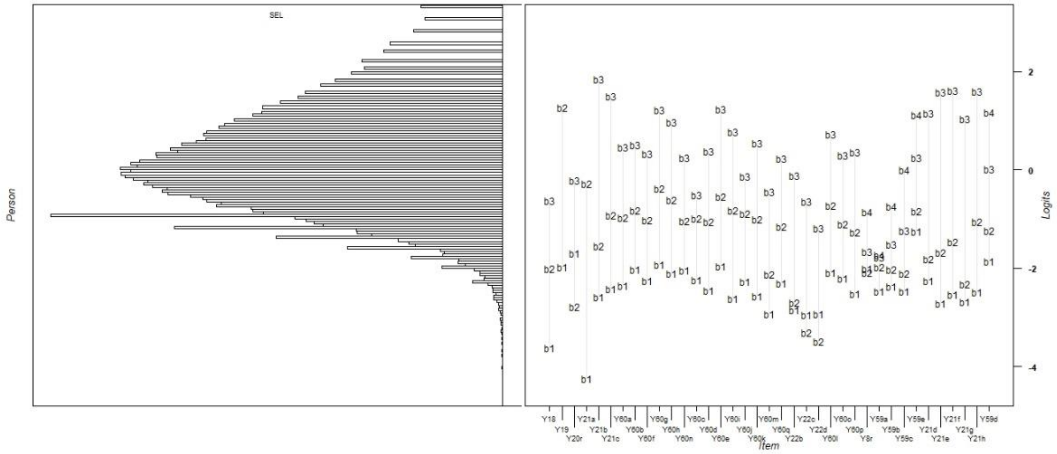
Table 4.9

Item Location Parameters for a Unidimensional Partial Credit Model

Item		Item transition locations			
	Label	δ_1	δ_2	δ_3	δ_4
1	Y18	-3.62	-2.00	-0.62	
2	Y19	-1.97	1.28		
3	Y20r	-1.69	-2.77	-0.21	
4	Y21a	-4.24	-0.28		
5	Y21b	-2.58	-1.55	1.85	
6	Y21c	-2.42	-0.91	1.51	
7	Y60a	-2.34	-0.96	0.47	
8	Y60b	-2.01	-0.82	0.52	
9	Y60f	-2.25	-1.01	0.34	
10	Y60g	-1.91	-0.38	1.22	
11	Y60h	-2.09	-0.60	0.98	
12	Y60n	-2.03	-1.03	0.25	
13	Y60c	-2.22	-0.98	-0.51	
14	Y60d	-2.44	-1.05	0.38	
15	Y60e	-1.95	-0.53	1.23	
16	Y60i	-2.61	-0.82	0.79	
17	Y60j	2.26	-0.88	-0.12	
18	Y60k	-2.56	-1.00	0.55	
19	Y60m	-2.92	-2.11	-0.43	
20	Y60q	-2.30	-1.16	0.24	
21	Y22b	-2.84	-2.70	-0.11	
22	Y22c	-2.94	-3.30	-.64	
23	Y22d	-2.92	-3.48	-1.19	
24	Y60l	-2.10	-0.72	0.73	
25	Y60o	-2.19	-1.10	0.31	
26	Y60p	-2.51	-1.27	0.38	
27	Y8r	-2.00	-2.10	-1.66	-0.84
28	Y59a	-2.46	-1.97	-1.78	-1.71
29	Y59b	-2.36	-2.01	-1.51	-0.74
30	Y59c	-2.47	-2.10	-1.23	0.00
31	Y59e	-1.25	-0.83	0.25	1.13
32	Y21d	-2.25	-1.80	1.17	
33	Y21e	-2.70	-1.67	1.59	
34	Y21f	-2.53	-1.46	1.62	
35	Y21g	-2.67	-2.32	1.04	
36	Y21h	-2.47	-1.04	1.60	
37	Y59d	-1.86	-1.22	0.02	1.17

Figure 4.8 displays the underlying continuum for the SEL construct. The range of the items are adequate for the range of the latent trait levels among the participants between -4 and 2 logits. More items are needed for latent trait levels above 2 logits. The SEL latent trait continuum seems to be nonsymmetrical where there is a ceiling a ceiling effect. There are probably students with higher true scores, however, the SEL measure lacks information (no items) that are sensitive to the higher end of the scale.

Figure 4.8
Item-person Map for the Unidimensional SEL Construct



The model-fit indices for the unidimensional IRT model are reported in Table 4.10. The model fit is poor with a very low CFI (CFI = .63) and TLI (TLI = .63). The RMSEA = .10 and SRMSR = .14, both are high values. The AIC and BIC values are very high, yet are lower than the two and six latent trait models.

Table 4.10

Model-fit Statistics for Unidimensional and Multidimensional Partial Credit Model

	1-factor	2-factor	6-factor
LL	-608054	-733202	-934049
EsPa	17	19	25
AIC	1216142	1466442	1868148
AICc	1216142	1466442	1868148
BIC	1216306	1466625	1868389
SABIC	1216252	1466565	1868309
M2	4502	717	22833
df	4	2	11
p-value	<.0001	<.0001	<.0001
RMSEA	.10	.06	.14
SRMSR	.14	.16	.09
CFI	.63	.96	.44
TLI	.54	.94	.39

Note. EsPa = Estimated number of parameters.

Multidimensional IRT models

Two multidimensional IRT models are reported. The two-dimensional latent trait model (Developmental Skills and Supports latent traits) item intercepts are reported in Table 4.11. The six-dimensional latent trait model (CtL, PI&O, SC, EM, FCS, and TSS) item intercepts are reported in Table 4.12.

Table 4.11

Item Intercept Parameters for a Two Dimensional Partial Credit Model

Item			Item intercepts			
	Trait	Label	γ_1	γ_2	γ_3	γ_4
1	Skills	Y18	3.69	5.70	6.32	
2	Skills	Y19	1.98	0.69		
3	Skills	Y20r	1.74	4.52	4.72	
4	Skills	Y21a	4.27	4.55		
5	Skills	Y21b	2.62	4.17	2.30	
6	Skills	Y21c	2.45	3.35	1.84	
7	Skills	Y60a	2.37	3.33	2.85	
8	Skills	Y60b	2.03	2.85	2.32	
9	Skills	Y60f	2.28	3.29	2.94	
10	Skills	Y60g	1.93	2.30	1.07	
11	Skills	Y60h	2.11	2.71	1.72	
12	Skills	Y60n	2.05	3.08	2.82	
13	Skills	Y60c	2.26	3.24	3.74	
14	Skills	Y60d	2.47	3.53	3.13	
15	Skills	Y60e	1.97	2.49	1.25	
16	Skills	Y60i	2.64	3.46	2.66	
17	Skills	Y60j	2.29	3.17	3.28	
18	Skills	Y60k	2.59	3.59	3.03	
19	Skills	Y60m	2.98	5.11	5.54	
20	Skills	Y60q	2.33	3.49	3.23	
21	Supports	Y22b	3.21	6.13	6.23	
22	Supports	Y22c	3.34	6.90	7.58	
23	Supports	Y22d	3.34	7.11	8.37	
24	Supports	Y60l	2.34	3.16	2.30	
25	Supports	Y60o	2.47	3.71	3.33	
26	Supports	Y60p	2.80	4.21	3.76	
27	Supports	Y8r	2.38	4.76	6.62	7.48
28	Supports	Y59a	2.85	5.16	7.18	8.96
29	Supports	Y59b	2.73	5.04	6.74	7.49
30	Supports	Y59c	2.83	5.20	6.58	6.52
31	Supports	Y59e	1.49	2.43	2.15	.813
32	Supports	Y21d	2.55	4.49	3.22	
33	Supports	Y21e	3.01	4.81	3.08	
34	Supports	Y21f	2.82	4.40	2.62	
35	Supports	Y21g	3.01	5.49	4.37	
36	Supports	Y21h	2.74	3.87	2.11	
37	Supports	Y59d	2.14	3.51	3.50	2.13

Note. Item discriminations are fixed at 1.00.

Skills = Developmental Skills, Supports = Developmental Supports.

Variance (Developmental Skills) = 1.21; Variance (Developmental Supports) = 1.68; Covariance (Developmental Skills, Developmental Supports) = 1.22.

Table 4.12

Item Intercept Parameters for a Six Dimensional Partial Credit Model

Item			Item intercepts			
	Trait	Label	γ_1	γ_2	γ_3	γ_4
1	CtL	Y18	2.16	4.04	4.59	
2	CtL	Y19	2.13	1.02		
3	CtL	Y20r	0.49	3.30	3.46	
4	CtL	Y21a	4.17	4.31		
5	CtL	Y21b	2.01	3.49	1.77	
6	CtL	Y21c	2.03	2.64	1.51	
7	PI&O	Y60a	1.74	2.40	2.26	
8	PI&O	Y60b	1.47	2.00	1.84	
9	PI&O	Y60f	1.62	2.34	2.30	
10	PI&O	Y60g	1.54	1.71	.95	
11	PI&O	Y60h	1.67	2.0	1.44	
12	PI&O	Y60n	1.39	2.15	2.16	
13	SC	Y60c	1.50	2.15	2.75	
14	SC	Y60d	1.80	2.55	2.46	
15	SC	Y60e	1.57	1.86	1.07	
16	SC	Y60i	2.10	2.60	2.19	
17	SC	Y60j	1.61	2.17	2.47	
18	SC	Y60k	1.95	2.64	2.42	
19	SC	Y60m	1.64	3.69	4.05	
20	SC	Y60q	1.61	2.49	2.50	
21	EM	Y22b	1.41	4.12	4.21	
22	EM	Y22c	1.22	4.49	4.98	
23	EM	Y22d	1.07	4.27	5.64	
24	EM	Y60l	1.59	2.03	1.70	
25	EM	Y60o	1.53	2.36	2.33	
26	EM	Y60p	1.78	2.78	3.76	
27	FCS	Y8r	0.47	1.52	3.01	3.93
28	FCS	Y59a	0.70	1.43	2.73	5.17
29	FCS	Y59b	0.79	1.84	3.17	3.96
30	FCS	Y59c	0.96	2.30	2.24	3.50
31	FCS	Y59e	1.49	0.50	0.97	0.82
32	TSS	Y21d	1.49	3.18	2.32	
33	TSS	Y21e	2.02	3.61	2.20	
34	TSS	Y21f	1.97	3.27	1.88	
35	TSS	Y21g	1.61	3.95	3.14	
36	TSS	Y21h	2.07	2.83	1.56	
37	TSS	Y59d	1.01	1.81	1.78	1.22

Note.

Commitment to learning (CtL), Positive Identity and Outlook (PI&O), Social Competence (SC), Empowerment (EM), Family/Community Supports (FCS) and Teacher/School Supports (TSS).

The latent trait covariances are reported in Table 4.13.

Table 4.13
Latent Trait Covariance for Six Dimensional Model

	CtL	PI&O	SC	EM	FCS
PI&O	.558				
SC	.546	.900			
EM	.662	.576	.561		
FCS	.646	.599	.584	.653	
TSS	.598	.530	.519	.587	.573

Note. Developmental Supports include Commitment to Learning (CtL), Positive Identity and Outlook (PI&O) and Social Competence (SC). Developmental supports include Empowerment (EM), Family/Community Supports (FCS) and Teacher/School Supports (TSS).

MI for Bifactor Model 2

Research question 2 tested for MI across eight ethnic groups to have evidence to compare the scores across the eight groups. MI for the six domain specific factors was conducted for bifactor model 2. The configural model fits worse than the fully saturated model ($\chi^2 = 441141$, $(df = 4708)$, p<.001, RMSEA = .081, CFI = .907). Although the model has an acceptable RMSEA (RMSEA close to .08) and adequate CFI (CFI > .90) the weak model fits worse than the fully saturated model ($\chi^2 = 437363$, $(df = 4918)$ p<.001, RMSEA = .079, CFI = .908). The chi-square of difference test between the configural and weak model had a p-value smaller than .05. This suggests there is no evidence of weak invariance for the bifactor model with a general factor and six specific domains ($\Delta\chi^2 = 418.82, (\Delta df = 210), p < .001, \alpha = .05, \Delta CFI = .001, \Delta RMSEA = -.002$). Instead of looking for partial MI in the bifactor model 2, a six factor (less complex) model and next best fitting model was tested for

MI. The MI models were conducted using the R software version 4.0.1 (R Core Team, 2020) and WLSMV estimator.

External Criterion Prediction

A structural linear model briefly examined the extent to which the six specific domain factors (CtL, PI&O, SC, EM, FCS, and TSS) predict Grades above and beyond the general factor (Developmental Skills and Supports). The matrix form for the model is expressed as:

Measurement model

$$Y = \Lambda_y \boldsymbol{\eta} + \boldsymbol{\epsilon}$$

Structural model

$$\eta_1 = \beta_{g1,38}\eta_{g1} + \beta_{s1,38}\eta_{s1} + \beta_{s2,38}\eta_{s2} + \beta_{s3,38}\eta_{s3} + \beta_{s4,38}\eta_{s4} + \beta_{s5,38}\eta_{s5} + \beta_{s6,38}\eta_{s6} + \zeta_1$$

The general factor and six domain specific factors provided an adequate fit: chi-square (622, N = 126812) = 523730.107, p < .001, CFI = .898, SRMR = .057, and RMSEA = .081 (Kline, 2015). The G factor ($\beta_{g1,1} = .298$, SE = .002 p < .001), CtL ($\beta_{s1,1} = .375$, SE = .003 p < .001), PI&O ($\beta_{s2,1} = .014$, SE = .003 p < .001), SC ($\beta_{s3,1} = .207$, SE = .003 p < .001), EM ($\beta_{s4,1} = .052$, SE = .003 p < .001), FCS ($\beta_{s5,1} = .034$, SE = .004 p < .001), and TSS ($\beta_{s6,1} = .072$, SE = .003 p < .001) predicted Grades. The domain-specific factors predicted Grades above and beyond the General factor.

Two-tier Random Intercept Model

Research question 3a was about the extent to which the lack of MI may be due to the idiosyncratic response style by using a two-tier model with a random intercept. The two-tier model with a random intercept basically was the bifactor model 2 (one general factor and six domain specific factors) with the random intercept to account for the variance due to the magnitude of the effect of individual idiosyncratic response style. However, the two-tier model with a random intercept did not converge in *FlexMIRT* 3.5. Although other less complex CFA models converged in *FlexMIRT* and *Mplus* 8.3, the two-tier models solely relied on the *FlexMIRT* software.

Gathering some evidence from the item loadings in the two-tier model with two primary factors and six specific factors, not all items load positively on the factors and there are five items with negative standard errors which may have contributed to the lack of the two-tier random intercept model converging.

More MI

Since the two-tier model with a random did not converge and no evidence of MI property was found in the bifactor model 2, research question 3b tested MI to the next best fitting more simple model. For the six-factor structure, the configural model fits worse than the fully saturated model ($\chi^2(4913) = 562959$, p<.001, RMSEA = .090, CFI = .881). Similarly, the weak model fits worse than the fully saturated model ($\chi^2(5130) = 527700$, p<.001, RMSEA = .085, CFI = .89) since the *p*-value is not above .05. The chi-square difference test resulted in the configural model having an

improvement in fit over the weak model since $\Delta\chi^2 = 429.1,\ (\Delta df = 217), p < .001, \alpha = .05, \Delta CFI = .01, \Delta RMSEA = -.005$). In other words, there is no evidence of weak MI for the six-factor model.

The next more simple model, the two-factor model was tested for MI. The two-factor structure, the configural model fits worse than the fully saturated model ($\chi^2(5024) = 859723$, p<.001, RMSEA = .110, CFI = .818). The same was true for the weak model where the weak model fits worse than the fully saturated model ($\chi^2(5269) = 797994$, p<.001, RMSEA = .103, CFI = .831) since the *p*-value is not above .05. The chi-square difference test resulted in the configural model having an improvement in fit over the weak model since $\Delta\chi^2 = -61729,\ (\Delta df = 245), p < .001, \alpha = .05, \Delta CFI = .012,\ \Delta RMSEA = -.007$). The two factor model does not hold the weak MI property.

The final and most simple (most restricted) one-factor model was tested for MI. The configural model fits worse than the fully saturated model ($\chi^2(5032) = 1047265$, p<.001, RMSEA = .121, CFI = .778). Also, the weak model fit worse than the fully saturated model ($\chi^2(5266) = 952721$, p<.001, RMSEA = .113, CFI = .798) since the *p*-value is not above .05. The chi-square difference test resulted in the configural model *not* having an improvement in fit over the weak model since $\Delta\chi^2 = 731,\ (\Delta df = 256), p > .05, \alpha = .05, \Delta CFI = .02,\ \Delta RMSEA =$.02). In other words, the weak model does not fit worse than the configural model. The result suggests the one factor model holds the MI property at a weak level. However, the one-factor

model did not the strong MI level. The weak property suggest that the SEL construct has similar meaning for the eight-ethnic community groups (Newson, 2015).

Chapter V: Discussion

SEL constructs have been welcomed by many organizations (largely supported by efforts from CASEL, 2015; RAND Education and Labor, 2020; Character Lab, 2017; Buros Center for Testing, 2020; & AIR, 2020a), the educational community (CORE Districts, 2018; MN Department of Education, 2020; & New York State Education Department, 2019) and researchers (MN Youth Development Research Group, 2020) with the promising studies that SEL leads to positive outcomes that may help in promoting healthy youth development, improving persistence, and closing achievement gaps—helping youth and adults thrive. In the course of this study, our world is going through the COVID-19 pandemic (Centers for Disease Control and Prevention, 2020); our Minnesota community as well as around the world is remembering George Floyd (CNN, 2020) who was killed by a white police officer (Kare11, 2020) and protests occurred around the world that brought more awareness to the systemic racism many people of color (especially our Black community) face among many other injustice issues. Our youth and especially our youth of color are facing real-world events that have disrupted all areas of their lives (e.g., online educational delivery during the stay-at-home order) bringing a lot of stress.

As mentioned in Chapter 1, there is no single definition of SEL or a streamlined method to measure SEL constructs. There was a sense of urgency to put forth a paradigm to measure SEL constructs, but now with challenging world events facing our youth, there is priority in adopting a paradigm in measuring SEL constructs. This is a priority as many organizations and school districts have put SEL surveys at

the forefront as a way to support students and teachers, as school districts are navigating through COVID-19 (AIR, 2020b; CASEL, 2020; New York State Education Department, 2020; & RAND, 2020). In this study I proposed a paradigm to measure SEL constructs that supply context and support score interpretation and use where score interpretation and use is meaningful, useful, appropriate, and fair.

Summary of the Study

In Chapter 1, a paradigm was introduced as a way to measure SEL constructs (Figure 1.1). The paradigm borrows from several psychometric researchers' validity work. First, the parts needed for SEL assessment (Rodriguez, 2020) are: Framework, Construct, Measures, and Items. The researcher chooses a framework (e.g., Personality Psychology – Big Five, 21st Century Skills) to put context around the SEL construct(s) such as self-efficacy, self-regulation, positive identity just to name a few. The construct is operationalized through measures (e.g., CORE measures) composed of items. Second, the framework qualitatively defines the construct and the measures and items, from which responses quantitatively define the construct through the measurement model.

Third, is the measurement model (Wilson, 2005) where the participant is assumed to have some amount of construct (e.g., emotional intelligence) that causes the responses to the items. We score the item responses (e.g., in an outcome space such as 0 = false and 1 = true) and use a psychometric model (e.g., IRT model) to infer about the person's position on the underlying construct. The better a construct is

qualitatively defined in a plausible framework, the better the construct can be quantitatively measured.

Fourth, validity evidence is collected at the measure level and applied to the construct (Messick, 1989). Forms of validity may include: Content evidence, internal structure, associations with other variables, response processes, measurement invariance, and fairness (Testing Standards, 2014). Lastly, the scores are reported at the construct level and need to have accumulating evidence to support interpretation and use. The IUA is the set of claim or proposed interpretations and uses of the scores (Kane, 2013). The more claims the IUA has, the more forms of validity evidence may be necessary.

In this study the PYD framework was used to qualitatively define the SEL construct. The PYD framework comes from positive psychology and focuses on the assets of youth. These assets are common to all youth. The SEL construct is operationalized by the measures composed of 37 item in the MSS. These rating-scale items were scored using psychometric models of various dimensions (e.g., bifactor model) that are all plausible given the PYD framework.

CFA Findings

Through research question 1a, I found the best fitting model to be the bifactor model 2 with a general factor and six domain specific factors. Not all items loaded well on their specified group factor. However, most items loaded high on the general factor, suggesting the specific domain factors may not explain above and beyond the general factor. Since the bifactor model 1 and bifactor model 2 had the same degrees

of freedom (df = 592) the models could not be compared via a chi-square difference test. Also, since the six-domain factor model had a better fit than the two-factor, the two-tier model was chosen to have six domains instead of two.

IRT Findings

Applying the paradigm, the Wright map displays the estimated location of the participants on the left-side of the map and the calibrated item locations on the right-side of the map (Wilson, 2005). The Wright maps for the six one-dimensional IRT models (CtL, PI&O, SC, EM, FCS, and TSS), one SEL dimensional model, and two-dimensional were displayed. The one-dimensional model for SEL had more items locations that are adequate for the range of the SEL latent trait level among the participants. Perhaps more items with an item location of 2 logits or above may help in filling out the small gaps on the SEL latent trait. Also, the SEL latent trait was approximately normally distributed.

MI for Bifactor Model 2

Research question 2 dealt with MI and briefly with external criterion prediction for the bifactor model 2 since this model had the best model-fit. The general and six specific domain factors in bifactor model 2 were positively associated with grades. Therefore, youth who report higher grades tend to have higher CtL, PI&O, SC, EM, FCS, TSS, as well as general SEL. Also, CtL, PI&O, SC, EM, FCS, and TSS predicted above and beyond the general SEL measure. This can possibly lead to be able to report a total score of a general SEL construct and subscores for CtL, PI&O,

SC, EM, FCS, and TSS constructs (Chen et. al, 2012). However, there is no evidence to compare these scores across ethnic/race community groups as the bifactor model 2 did not have the MI property.

Two-tier Random Intercept Model

Research question 3a aimed at ruling out if the lack of MI for the bifactor model 2 was due to the effect of individual idiosyncratic response styles since the student ethnic/racial cultural background may affect their responses to the items. Unfortunately, after 80 hours of trying to have the two-tier model converge in *FlexMIRT 3.5*, it seems that this model needs a simulation study specifically for polytomous SEL items. Most of the research with this model is geared towards psychological constructs aimed for adults and not for youth.

More MI

Research question 3b tested other models for MI since the bifactor model 2 did not have evidence to support MI. The six, two and one factor models were tested for MI. The models were from more complex to least complex. At the end the one factor model had weak MI but not strong MI. This implies the factor loadings across the eight ethnic/race community groups can be compared but not their item means or factor mean scores.

Limitations of the Study and Future Research

There are several limitations to this study and a few will be briefly shared. The study was based on secondary data-analysis. Not all 37 items follow the wording advised by Haladyna and Rodriguez (2013), in terms of optimal survey item design. Also, the lack of MI could not be ruled due to idiosyncrastic person responses. Lastly, the Latino youth group was one group even though Latino youth come from various countries (or parents are from different countries), have different lengths of living in the US (e.g., first generation vs. fourth generation), and also come from different within-ethnic and racial backgrounds (e.g., indigenous heritage, Afro-Caribbean heritage).

There are endless future research opportunities in measuring SEL constructs; however, a few are mentioned here. First, there needs to be more studies of the item wording in SEL measures. As seen in achievement testing, the wording of the item is crucial to measurement quality. Secondly, there needs to more studies to investigate the MI property for the proposed groups that are being investigated. Many of the studies today, compare groups with a model that has not been tested for MI. And, it is possible that many of the SEL constructs being measured are poorly defined for multiracial and multiethnic populations. Lastly, more research needs to be done on SEL constructs with IRT models to investigate whether the latent trait continuum can be improved through certain items. Granted, it is difficult to write items in SEL that cover the full range of any given construct, such that measurement information is obtained across the entire trait continuum.

Conclusion

In conclusion, a paradigm was proposed in this study as a model of how to measure SEL constructs and evaluate measurement quality. Following this paradigm, the bifactor model 2 was the best model fit yet did not have the MI property to compare across eight ethnic/racial groups. The one-factor model had weak MI evidence to compare item loadings across the ethnic groups. The bifactor model, unfortunately, cannot be used to compare the mean factor scores across the eight groups. It is necessary to follow a paradigm that will lead to meaningful, useful, appropriate, and fair score interpretation and use. The SEL community of qualitative and quantitative researchers need to work together to properly define SEL constructs in a qualitative and quantitative manner.

References

American Institutes for Research (AIR). (2007). *Social Emotional Learning.* https://www.air.org/topic/education/social-and-emotional-learning

American Institutes for Research (AIR). (2020, July 16). *AIR's COVID-19 Response and Resources*. https://www.air.org/resource/air-s-covid-19-response-and-resources

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Agresti, A. (2013). Categorical data analysis (3rd ed., Wiley series in probability and statistics). Wiley-Interscience.

Albano, A. (2020). *Introduction to educational and psychological measurement using R.* https://thetaminusb.com/intro-measurement-r/index.html

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Benson, P. L. (2002). Adolescent development in social and community context: A program of research. *New Directions for Youth Development*, *95*, 123-47.

Benson, P.L., Scales, P.C., Hamilton, S.F., & Sesma, A. (2006). *Positive youth development: Theory, research, and applications*. In W. Damon & R.M. Lerner (Eds.), Handbook of Child Psychology: *Vol. 1* (6th ed., pp. 894-941). John Wiley & Sons.

Berg, J., Osher, D., Same, M. R., Nolan, E., Benson, D., & Jacobs, N. (2017). *Identifying, defining, and measuring social and emotional competencies*. American Institutes for Research.

Berman, S., Chaffee, S., & Sarmiento, J. (2018). *The practice base for how we learn: Supporting students' social, emotional, and academic development*. The Aspen Institute.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores.* Addison-Wesley.

Bock, R. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*(4), 21-33.

Bock, R., & Aitkin, D. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

Bollen, K. (1989). Structural equations with latent variables (Wiley series in probability and mathematical statistics. Applied probability and statistics). Wiley.

Brennan, R., & National Council on Measurement in Education. (2006). *Educational measurement* (4th ed., American Council on Education/Praeger series on higher education). Praeger.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*(4), 581-612.

Cai, L. (2019). flexMIRT: A numerical engine for flexible multilevel multidimensional item analysis and test scoring (Version 3.5) [Computer software]. Vector Psychometric Group.

Cai, L., Yang, J., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*(3), 221-248.

Care, E., Griffin P., & Wilson, M. (2018). Assessment and teaching of 21st century skills: Research and applications (Educational assessment in an information age). Springer.

CASEL. (2015). *CASEL Guide: Effective social and emotional learning programs - middle and high school edition*. http://secondaryguide.casel.org/casel-secondary-guide.pdf

CASEL. (2020). *CASEL cares initiative; Connecting the SEL community.* https://casel.org/resources-covid/

Centers for Disease Control and Prevention. (June, 2020). *COVID-19 Mathematical Modeling.* https://www.cdc.gov/coronavirus/2019-ncov/covid-data/mathematical-modeling.html

Chang, Y. (2015). *A restricted bi-factor model of subdomain relative strengths and weaknesses* (Publication No. 3733255) [Doctoral dissertation, University of Minnesota]. ProQuest Dissertations and Theses Global.

Character Lab. (2017). *Character strengths Playbooks.* https://characterlab.org/playbooks/

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of personality and social psychology*, *95*(5), 1005.

Chen, F., Hayes, A., Carver, C., Laurenceau, J., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, *80*(1), 219-251.

Chen, F., & Zhang, Z. (2018). Bifactor models in psychometric test development. In Irwing, P., Booth, T., and Hughes, D. (Eds.), *The Wiley handbook of psychometric testing : A multidisciplinary reference on survey, scale and test development*. (pp. 325-345). John Wiley & Sons.

Chernyshenko, O., Kankaraš, M., & Drasgow. F. (2018). *Social and emotional skills for student success and well-being: Conceptual framework for the OECD study on social and emotional skills*. OECD Education Working Papers, No. 173. https://www.oecd-ilibrary.org/education/social-and-emotional-skills-for-student-success-and-well-being_db1d8e59-en

Cizek, G. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31-43.

CNN. (June 3, 2020). *This is how loved ones want us to remember George Floyd.* https://www.cnn.com/2020/05/27/us/george-floyd-trnd/index.html

CNN. (June 13, 2020). *Protests across the globe after George Floyd's death.* https://www.cnn.com/2020/06/06/world/gallery/intl-george-floyd-protests/index.html

Costa, P.T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI) professional manual.* Psychological Assessment Resources.

Cox, J., Foster, B., & Bamat., D. (2019). *A review of instruments for measuring social and emotional learning skills among secondary school students*. Educational Development Center, Inc., Institute of Education Sciences U.S. Department of Education.

CORE Districts. (October, 2018). *Incorporating social-emotional learning into school accountability.* https://edsource.org/2015/incorporating-social-emotional-learning-into-school-accountability-bookman-commentary/88989

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Cross-agency project team. (2018). Social and emotional learning in California: A guide to resources. *California Department of Education.*

de Ayala, R. (2009). *The theory and practice of item response theory* (Methodology in the social sciences). New York: Guilford Press.

De Fruyt, F., Wille, B., & John, O. (2015). Employability in the 21st century: Complex (interactive) problem solving and other essential skills. *Industrial and Organizational Psychology-Perspectives on Science and Practices, 8*(2), 276-U189.

Domitrovich, C., Durlak, J., Staley, K., & Weissberg, R. (2017). Social-emotional competence: An essential factor for promoting positive adjustment and reducing risk in school hildren. *Child Development, 88*(2), 408-416.

Durlak, J., Weissberg, R., Dymnicki, A., Taylor, R., & Schellinger, K. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82*(1), 405-432.

Educational Reform. (2016, August 25). *21st Century Skills.* https://www.edglossary.org/21st-century-skills/

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists* (Multivariate applications book series). L. Erlbaum Associates.

Farrington, C.A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T.S., Johnson, D.W., & Beechum, N. O. (2012). *Teaching adolescents to become learners: The role of noncognitive factors in shaping school performance--A critical literature review*. Chicago, IL: University of Chicago, Consortium on Chicago School Research.

Flora, D., & Curran, P. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466-491.

Gabrieli, C., Ansel, D., & Krachman, S. B. (2015). Ready to be counted: The research case for education policy action on non-cognitive skills. *Transforming Education*.

Gehlbach, H., & Hough, H. J. (2018). Measuring social emotional learning through student surveys in the CORE Districts: A pragmatic approach to validity and reliability. *Policy Analysis for California Education, PACE*.

Gibbons, R., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423-436.

Griffin, P., McGaw, B., Care, E. (2012). *Assessment and teaching of 21st century skills*. Springer.

Gorges, J., Koch, T., Maehler, D. B., & Offerhaus, J. (2017). Same but different? Measurement invariance of the PIAAC motivation-to-learn scale across key socio-demographic groups. *Large-scale Assessments in Education, 5*(1), 13.

Gustafsson, J., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28*(4), 407-434.

Haladyna, T., & Rodriguez, M. (2013). *Developing and validating test items*. Routledge.

Hambleton, R., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of item response theory* (Measurement methods for the social sciences series). Sage Publications.

Hamilton, L.S., Stecher, B.M., Schweig, J., & Baker G. (2018). *RAND Education Assessment Finder*. https://www.rand.org/pubs/tools/TL308.html

Hull, J., Lehn, D., & Tedlie, J. (1991). A General Approach to Testing Multifaceted Personality Constructs. *Journal of Personality and Social Psychology, 61*(6), 932-945.

Jones, S. (2018). *The Taxonomy Project*. EASEL Lab, Harvard Graduate School of Education. https://easel.gse.harvard.edu/taxonomy-project

Jöreskog, K.G., & Sörbom D. (1989). LISREL 7: A guide to program and applications (2nd ed.). Chicago, IL: SPSS.

Kare11. (May 28, 2020). *Transcript of 911 call on George Floyd released.* https://www.kare11.com/article/news/local/george-floyd/transcript-of-911-call-on-george-floyd-is-released/89-34f18837-3b09-421b-b3db-e2c0f5dfa6fa

Kamata, A., Bauer, D.J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling. 15*(1), 136-153.

Kane, M.T. (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kane, M.T. (2013b). The argument-based approach to validation. *School Psychology Review, 42*(2), 448-457.

Kendziora, K., & Yoder, N. (2016). When districts support and integrate social and emotional learning (SEL): Findings from an ongoing evaluation of districtwide implementation of SEL. *Education Policy Center at American Institutes for Research*.

Kline, R.B. (2016). *Principles and practice of structural equation modeling* (fourth edition). The Guildford Press.

Knight, G., Roosa, M., Calderon-Tena, C., & Gonzales, N. (2009). Methodological issues in research on Latino populations. In F. A. Villarruel, G. Carlo, J.M Grau, M. Azmitia, N, J. Cabrera, & T. J. Chahin (Eds.), *Handbook of U.S. Latino psychology: Developmental and community-based perspectives* (pp.45-62). SAGE.

Kyllonen, P., & Zu, J. (2019, April 4). *Measuring social, emotional, and self-management skills for schools and the workplace.* [Pre-conference training session]. The annual meeting of the National Council on Measurement in Education, Toronto, Canada.

Lerner, R. M, Lerner, J.V, Almerigi, J.B., Theokas, C.P., Gestsdottir, S., & Von Eye, A. (2005). Positive youth development, participation in community youth development programs, and community contributions of fifth-grade adolescents: Findings from the first wave of the 4-H study of positive youth development. *The Journal of Early Adolescence, 25*(1), 17-71.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*, 5-55.

Liu, Y., Millsap, R., West, S., Tein, J., Tanaka, R., & Grimm, K. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods, 22*(3), 486-506.

Masters, G. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

Maydeu-Olivares, A., & Coffman, D. (2006). Random intercept item factor analysis. *Psychological Methods, 11*(4), 344-362.

McKinley, R., & Reckase, M. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research Report No. ONR83-2). American College Testing Program.

Millsap, R. (2011). *Statistical approaches to measurement invariance.* Routledge.

MN Department of Education. (2018). *Minnesota Student Survey*. Roseville, MN: Author. Retrieved from https://education.mn.gov/MDE/dse/health/mss/

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.

Meyer, R., Wang, C., & Rice, A.B. (2018). *Measuring students' social-emotional learning among California's CORE districts: An IRT modeling approach*. Policy Analysis for California Education, Stanford Graduate School of Education. Retrieved from http://www.edpolicyinca.org/publications/sel-measurement

MN Department of Education. *SEL Implementation Guidance*. Author. https://education.mn.gov/MDE/dse/safe/social/imp/

Minnesota Youth Development Research Group. (2020). *Research group publications*. University of Minnesota. https://conservancy.umn.edu/handle/11299/194886

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

Muthén, L.K., & Muthén, B.O. (2019). Mplus (Version 8.3). [Software program]. Los Angeles, CA: Authors.

New York State Education Department. (2018). *Social Emotional Learning Activities and Teaching Practices.* Author. http://www.p12.nysed.gov/sss/SELCrosswalks.html

New York State Education Department. (2019) *Social Emotional Learning Resources.* (2020). Author. http://www.nysed.gov/edtech/educator-resources

Panayiotou, M., Humphrey, N., & Wigelsworth, M. (2019). An empirical basis for linking social and emotional learning to academic performance. *Contemporary Educational Psychology*, 56, 193-204.

RAND Education and Labor. *About the SEL Center*. https://www.rand.org/education-and-labor/centers/sel/about.html

RAND. (May 26, 2020). *COVID-19 and the State of K-12 schools.* https://www.rand.org/pubs/research_reports/RRA168-1.html

Reckase, M. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*(4), 401-412.

Reckase, M. (2009). *Multidimensional Item Response Theory* (Statistics for Social and Behavioral Sciences book series). Springer.

Reckase, M. (2017). *A tale of two models: Sources of confusion in achievement testing* (Research Report No. RR-17-44). Educational Testing Service.

Reise, S.P., & Revicki, D.A. (2015). *Handbook of item response theory modeling: Applications to typical performance assessment (Multivariate applications book series)*. Routledge.

Restorative Practices Working Group. (2014). Restorative practices: Fostering healthy relationships and promoting positive discipline in schools. *Opportunity to Learn Campaign.*

Rhemtulla, M., Brosseau-Liard, P.E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354-373.

Rodriguez, M. C., (2017). *Technical report on developmental skills, supports, & challenges: 2013-2016 Minnesota Student Survey*. Minneapolis, MN: University of Minnesota. https://conservancy.umn.edu/handle/11299/194886

Rodriguez, M. C. (2018a, July 11). A psychometric perspective on SEL assessment. [Web blog post]. *Measuring SEL: Using data to inspire practice*. CASEL Measuring SEL Network. http://measuringsel.casel.org/psychometric-perspective-sel-assessment/

Rodriguez, M. C. (2018b). *Assessment of social and emotional learning*. [Unpublished manuscript]. Minneapolis, MN: Educational Psychology Department, University of Minnesota.

Rosen, J., Glennie., E., Dalton, B., Lennon, J., & Bozick, R. (2010). *Noncognitive skills in the classroom: New perspectives on educational research*. RTI Press/RTI International.

Ross, K., & Tolan, P. (2018). Social and emotional learning in adolescence: Testing the CASEL model in a normative sample. *The Journal of Early Adolescence*, *38*(8), 1170-1199.

R Core Team (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://wwwR-project.org/

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*, No. 17.

Buros Center for Testing. (2020). *Social and emotional learning assessment technical guidebook.* Lincoln, NE: Author, University of Lincoln. https://buros.org/sel-assessment-technical-guidebook

Soto, C., & John, O. (2017). The next big five inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117-143.

Takane, Y., & Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393-408.

Taylor, J., Buckley, K., Hamilton, L. S., Stecher, B. M., Read, L., & Schweig, J. (2018). Choosing and using SEL competency assessments: What schools and districts need to know. *RAND Corporation.*

Search Institute. (2005). *Developmental Assets Profile user manual.* Minneapolis, MN: Author.

Thurstone, L., & Chave, E. (1929). *The measurement of attitude: A psychophysical method and some experiments with a scale for measuring attitude toward the church*. University of Chicago Press.

van der Linden, W. J. (Ed.). (2017). *Handbook of item response theory, Volume one: Models.* Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences.

Wang, J., Wang, X. (2012). *Structural equation modeling with Mplus: Methods and applications* (Wiley series in probability and statistics). Wiley/Higher Education Press.

Wilson, M. (2005). *Constructing measures an item response modeling approach*. Lawrence Erlbaum Associates.

Yao, L., & Schwarz, R. (2006). A Multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement, 30*(6), 469-492.

Yoder, N. (2014). Teaching the whole child: Instructional practices that support social-emotional learning in three teacher evaluation frameworks. Research-to-Practice Brief. *Center on Great Teachers and Leaders*.